
Reference
Not to be taken
from this library

The New Encyclopædia Britannica

Volume 27

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Jorge Aguilar-Cauz, President

Chicago
London/New Delhi/Paris/Seoul
Sydney/Taipei/Tokyo

First Edition 1768–1771
Second Edition 1777–1784
Third Edition 1788–1797
Supplement 1801
Fourth Edition 1801–1809
Fifth Edition 1815
Sixth Edition 1820–1823
Supplement 1815–1824
Seventh Edition 1830–1842
Eighth Edition 1852–1860
Ninth Edition 1875–1889
Tenth Edition 1902–1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003, 2005
By Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Control Number: 2004110413
International Standard Book Number: 1-59339-236-2

Britannica may be accessed at http://www.britannica.com on the Internet.

CONTENTS

1 San Francisco
6 São Paulo
10 Scandinavian Literature
22 Classical Scholarship
32 The History of Science
42A Science Fiction
43 The Art of Sculpture
60 The History of Western Sculpture
115 Sensory Reception
223 Seoul
226 Set Theory
233 Sex and Sexuality
253 Shakespeare
273 Shanghai
278 Shintō
284 Sikhism
288 Slavery
301 Sleep and Dreams
312 Adam Smith
316 The Social Sciences
365 Social Structure and Change
372 Social Welfare
393 Modern Socio-Economic Doctrines and Reform Movements
438 Socrates
441 Soils
452 The Solar System
556 Sound
578 South America
634 South Asian Arts
711 Southeast Asia
795 Southeast Asian Arts
838 Southern Africa

CONTENTS

San Francisco

San Francisco holds a secure place in the United States' romantic dream of itself—a cool, elegant, handsome, worldly seaport whose steep streets offer breathtaking views of one of the world's greatest bays. According to the dream, San Franciscans are sophisticates whose lives hold full measure of such civilized pleasures as music, art, and good food. To San Franciscans their city is a magical place, almost an island, saved by its location and history from the sprawl and monotony that afflicts so much of California.

Since World War II, however, San Francisco has had to face the stark realities of urban life: congestion, air and water pollution, violence and vandalism, and the general decay of the inner city. San Francisco's makeup has been changing as families, mainly white and middle-class, have moved to its suburbs, leaving the city to a population that, viewed statistically, tends to be older and to have fewer married people. Now almost one of every two San Franciscans is "nonwhite"—African American, East Asian, Filipino, Samoan, Vietnamese, Latin American, or Native American. Their dreams increasingly demand a realization that has little to do with the romantic dream of San Francisco. But both the dreams and the realities are important, for they are interwoven in the fabric of the city.

Although San Franciscans complain of the congestion, homelessness, and high cost of living that plague the city and talk endlessly of the good old days, the majority still think of San Francisco the way poet George Sterling did, as "the cool grey city of love," America's most attractive, colourful, and distinctive place to live.

This article is divided into the following sections:

Physical and human geography 1
- The landscape 1
  - The city site
  - Climate
  - The city layout
- The people 2
- The economy 3
  - The port
  - Industry and tourism
  - Finance
  - Transportation
- Administration and social conditions 3
  - Government
  - Public utilities
  - Education
- Cultural life 4
  - The arts
  - Cultural institutions
  - Popular culture

History 4
- Exploration and early settlement 4
- The growth of the metropolis 5
- The 20th century 5
  - Expansion during the world wars
  - From peace to protest
  - The late 20th century

Bibliography 5

Physical and human geography

THE LANDSCAPE

The city site. Hilly and roughly square, San Francisco occupies the northern tip of a peninsula. To its south are the bedroom suburbs of San Mateo county, to the east and northeast is the bay, and to the west and northwest lies the Pacific Ocean.

The most prominent of San Francisco's hills are Twin Peaks, Mount Davidson, and Mount Sutro, all of which exceed 900 feet (270 metres) in elevation. The best known are Nob Hill, where the wealthy "nobs" (nabobs) built extravagant mansions in the 1870s, and Telegraph Hill, which once looked down on the Barbary Coast, a neighbourhood alive with gaudy wickedness. As a result of the pioneer planners' prejudice in favour of a squared-off grid, the downtown streets march intrepidly up precipitous slopes, terrifying newly arrived drivers, and providing splendid views of the bay.

The city by the bay

San Francisco Bay is a drowned river valley, submerged during the melting of the last glacial ice sheet. Enthusiastic and profitable filling of the tidelands has reduced its area at mean high tide from about 700 square miles (1,800 square km) in 1880 to a mere 435 square miles (1,125 square km). More than half of the bay is still fillable, but in 1965 the state legislature created the Bay Conservation and Development Commission to control further landfill projects. At its widest extent the bay measures 13 miles (21 km) across; its deepest point, 357 feet (109 metres), is in the Golden Gate, a narrow channel between the peninsula and Marin county to the north that connects the bay to the Pacific. The maximum daily flow of water through the Golden Gate into the Pacific is seven times the flow of the Mississippi River at its mouth.

Within the portion of San Francisco Bay lying inside the city limits are the natural islands of Alcatraz and Yerba Buena and man-made Treasure Island, created for a world's fair in 1939 and later turned into a naval base (1941–93). Alcatraz (Spanish: "Pelican") was from 1934 to 1963 the most notorious maximum-security, "escape-proof" prison in the United States. The island became part of Golden Gate National Recreation Area in 1972 and has become a popular tourist attraction.

Climate. Winter in San Francisco is rainy and mild, spring sunny and temperate, summer foggy and cool, and autumn sunny and warm. The average minimum temperature is 51° F (11° C), and the average maximum is 63° F (17° C). The mean rainfall, almost all of which occurs between November and April, is about 21 inches (533 mm). The most characteristic feature of the weather, however, is the summer fog, which lies low over the city until midday, creating consternation among shivering tourists. This fog is a phenomenon of temperature contrasts, created when warm, moist ocean air comes in contact with cold water welling up from the ocean bottom along the coast.

The fog

The city layout. The central business district, the financial district, North Beach, and Chinatown occupy the site of the gold-rush city, which subsequently was expanded by progressive fillings along the waterfront. The remnants of many ships that were deserted in 1849 now lie under office buildings several blocks inland. To the west, at the approach to the Golden Gate Bridge, lies the Presidio, a two-century-old military installation that became part of Golden Gate National Recreation Area in 1994; it is remarkable for its parklike lawns and wind-sculptured stands of trees. South of the Presidio is Golden Gate Park, reclaimed from a onetime sandy desert. The rest of San Francisco is largely composed of residential neighbourhoods, from Pacific Heights, in which the old, wealthy families reside, to Hunter's Point, which is predominantly an African American community. Most are filled with flower-decked houses of pastel stucco and "painted ladies"—frame structures with abundant and often elegant architectural detailing, intricately coloured.

A great change, which has been described as the Manhattanization of San Francisco, became apparent after the late

1960s, and it has been both welcomed and resisted. In the financial district, in particular, one tall building after another has been constructed in a city in which, for generations, few structures were higher than 20 stories. Among the modern skyscrapers are Bank of America, the Transamerica Pyramid (which rises to an elongated point), and the Park Hyatt. The Hyatt Regency, known for its spectacular 20-story hanging garden, is part of the massive Embarcadero Center complex—designed by John Portman in the 1970s—which encompasses six city blocks and houses numerous shops, hotels, and restaurants.

Another concern is one that San Francisco shares with few other U.S. cities—destruction by earthquake. Severe quakes have been felt in 1868, 1898, 1900, 1906, and 1989. But it was the 1906 earthquake that did the most damage and that has become identified with the city. A little after 5:00 AM on April 18 the entire city began to tremble and shake. There was a terrible noise, "like the roar of 10,000 lions." Cable cars jerked to a stop and the $7 million City Hall crumbled like a movie set. The glass roof over the Palace Hotel court splintered and rained down shards.

Earthquakes

That quake was followed by a massive fire that destroyed the centre of town and burned for four days, until the smouldering ashes were wetted down by rain. Starting in the business section near Montgomery Street and the South of Market district, the inferno swept toward Russian Hill, Chinatown, North Beach, and Telegraph Hill, where Italians poured wine on the flames to save their houses. Gone were 4 square miles (10 square km), making up 512 blocks in the centre of town, along with 28,000 buildings and a total property value of about $350 million. Approximately 700 people died, and 250,000 were left homeless. Survivors camped in Golden Gate Park. Since the 1906 earthquake, seismologists and engineers have warned that it could happen again. Several relatively strong earthquakes (measuring more than 5.0 on the Richter scale) have since then caused little damage. But the quake on Oct. 17, 1989, which measured 7.1 on the Richter scale, killed more than 60 people and caused severe damage to the Marina District and to some freeways and even more devastation to surrounding areas. Modern office towers were largely unaffected, indicating that new building methods may provide some protection.

THE PEOPLE

The pattern of immigration into San Francisco during the latter half of the 19th century was significantly different from that of anywhere else in the United States. The waves of newcomers included not only native-born Americans moving west but also Europeans arriving directly by ship. The demography of the gold-rush city was summed up concisely by a real-estate firm that advertised it could "transact business in the English, French, German, Spanish and Italian languages." San Francisco remains one of the most Mediterranean of American cities—New Orleans is another—and Italians are still the dominant European minority, followed by Germans, Irish, and British.

Ethnic mix

Jewish immigrants from Europe arrived in the city even before the gold seekers of 1849, and much credit for San Francisco's culture must be given to them. They founded libraries, symphonies, and theatres and gave the city its first aura of sophistication.

Before World War II about 20,000 African Americans lived in the entire Bay Area, about 4,000 of them in San Francisco. The increase in the black population during the next 30 years was set in motion by the war, which brought at least a half million workers to the Bay Area's shipyards and other industries. Among them were tens of thousands from the South, who settled mainly in San Francisco, Oakland, and Richmond. In San Francisco they moved into the old Carpenter Gothic houses in the blocks around Fillmore Street, vacated when the Japanese who had lived there were driven into wartime internment camps. By the 1980s, the character of the district shifted again, as gentrification caused rents to skyrocket. Poorer African American residents were forced into slum housing in the city's already crowded southeastern sector. An increasing number of African Americans have become prominent in the city's life—Willie Brown was elected mayor in 1995 and reelected in 1999—and many others have won elective office.

Chinatown, the best-known Chinese community in the United States, is also probably the least understood minority community in the city. The colourful shops and restaurants of Grant Avenue mask a slum of crowded tenements and sweatshops that has the highest population density in a densely populated city. Chinese residents have increasingly moved into North Beach, hitherto predominantly Italian, onto Russian Hill, or into the middle-class neighbourhoods of the Richmond district north of Golden Gate Park, where some of the city's most popular Chinese restaurants and bakeries are found on Clement Street. Many who reside in Chinatown are more recent immigrants, particularly from Hong Kong.

The Japanese community of San Francisco was wiped out at a single stroke by the infamous Executive Order 9066 of 1942, which sent them, foreign-born and citizen alike, into "relocation centres." The present centre of the Japanese community is Japantown (Nihonmachi), a few blocks east of Fillmore Street, now an ambitious commercial and cultural centre. Though the rising generation of Japanese Americans go to Japantown as visitors, bound for church services, social or cultural events (such as the annual cher-

Colour Library International

The Coit Memorial Tower on Telegraph Hill, San Francisco; at left is Alcatraz Island in San Francisco Bay.

ry blossom festival), or shops selling imported goods, their own roots are elsewhere.

The Spanish-speaking population is the second-largest ethnic minority (after the Chinese). Before World War II the Mission District, named for the Mission Dolores, was principally Irish. The Irish were largely replaced by Spanish-speaking Latin American immigrants, mainly from Central America and Mexico. Living among them are pockets of Native Americans and Samoans.

The Filipino community has grown remarkably since World War II and has spread to all areas of the city, especially the South of Market area. Though not as numerous as in Southern California, the Vietnamese, Cambodian, and Laotian communities have grown considerably since the 1980s, conflicting with blacks and Hispanics over low-income housing. These groups created a proliferation of Asian restaurants in the troubled Tenderloin area between the Civic Center and Union Square.

The gay community

San Franciscans have historically considered their city to be laissez-faire and open-minded, which is probably why homosexuals have felt comfortable there. The affluent Castro district (technically Eureka Valley near Twin Peaks) has attracted gays and lesbians from throughout the country, becoming perhaps the most famous gay neighbourhood in the world. It is said that no local politician can win an election without the gay community's vote.

THE ECONOMY

The gold rush (1848–49) established San Francisco as the premier city of the West. It is still a great port, the financial and administrative capital of the West, and a substantial centre for commerce and manufacturing.

A large portion of the city's employed work in the area of finance. Other leading areas of employment include business services (personnel supply, building maintenance, security, computers and data processing, and advertising), retail trade, the tourist and convention industry, and professional services. Many companies, such as Levi Strauss & Co., producer of one of San Francisco's most famous products, blue jeans, have located their national headquarters in the Bay Area.

The port. From its beginnings as a port of call in the hide-and-tallow trade and, later, as the home port of the Pacific whale fishery, San Francisco has been acutely conscious of the importance of shipping. In the 19th century ships stopped there from their trip around Cape Horn or the Isthmus of Panama; after 1914 cargo and passenger vessels arrived from the East by way of the Panama Canal. In 1867 the Pacific Mail Steamship Company opened the first transpacific service, sailing from San Francisco to Yokohama (Japan) and Hong Kong. Imports and exports now passing through the San Francisco Customs District make the combined ports of San Francisco Bay—San Francisco, Oakland, Alameda, Sacramento, and Stockton—one of the most active international ports in the country.

Industry and tourism. Manufacturing is the main source of income in the Bay Area. In San Francisco, in which manufacturing is less important, the principal industries are apparel and other textile products, food processing, and shipbuilding, while the aerospace and electronics industries are strong in the cities of the peninsula. Of note is Silicon Valley, a region just south of the bay that is the heart of the nation's computer industry.

Tourist attractions

Tourism is a major source of income. The bridges, Coit Tower, the museums, the restaurants, Chinatown, North Beach, the Victorian mansions, crooked Lombard Street, and the dazzling Fairmont Hotel are major attractions; Fisherman's Wharf, however, is the most popular. Families browse the area, watching fishermen prepare the crab catch and mend their nets amid dozens of souvenir shops, street entertainers, restaurants, and bakeries selling one of the city's specialties, sourdough bread. The Powell–Hyde Street cable car is a popular route to the wharf.

San Francisco's waterfront offers whale-watching excursions and provides a boat tour from the wharf to Alcatraz Island. It is home to Ghirardelli Square, the onetime chocolate factory; the Cannery, built for the California Fruit Canners Association (now Del Monte Corporation) in 1907 and now a marketplace; Pier 39, reconstructed using timbers from old ships to create a New England look, home to shops and eateries and one of the best seal-watching spots on the coast; and the Anchorage, which has a mini-amphitheatre. Nearby is the Marina District, formerly known as Harbor View when it was the scene of the 1915 Panama-Pacific International Exposition.

Finance. A financial centre since the first pinch of gold dust was exchanged for cash, San Francisco is the seat of the Pacific Stock Exchange as well as the headquarters of many banks, among them the Bank of America and Wells Fargo. Though there are no native, independent banks headquartered in San Francisco, the city still ranks among the nation's largest investment banking centres.

Transportation. Periodic smog, produced mainly by the automobiles in the area, is a serious concern. Traffic is also a problem, as travel from the East Bay cities of Oakland and Berkeley and from Marin county to the north is confined to two great but overburdened bridges. The world's longest high-level steel bridge, the San Francisco–Oakland Bay Bridge, is 4.5 miles (7.2 km) long; it was completed in 1936 and consists of two back-to-back suspension bridges, a connecting tunnel on Yerba Buena Island, five truss spans, and a cantilever span. The Golden Gate Bridge, leading north to Marin county, was completed in 1937. It is a pure suspension bridge with a 4,200-foot (1,280-metre) centre span; the spectacular clear span was the longest in the world until 1964 when New York City's Verrazano-Narrows Bridge opened.

Bridges

Until the ferries were doomed by the bridges, San Francisco was served by a great network of ferry routes, whose splendid vessels were said to deliver more passengers to the Ferry Building at the foot of Market Street than arrived at any other transportation depot except Charing Cross railway station in London. Only after the bridges began to choke with traffic did the ferries return, on a smaller scale, between San Francisco and Marin county.

A much greater undertaking was the interurban rapid-transit system known as BART (Bay Area Rapid Transit), which began operating in 1972. With service between San Francisco and the East Bay communities through an underwater tube more than 3.6 miles (5.8 km) long, BART was the first system of its sort—part subway and part elevated—to be built in half a century. These comfortable, computerized automatic trains run at speeds as high as 80 miles (130 km) per hour.

San Francisco, situated at the head of a peninsula, has always been a dead end for rail traffic. Beginning with the arrival of the first westbound train over the tracks of the Central Pacific on Sept. 6, 1869, transcontinental trains began discharging their passengers in Oakland, where ferries or buses carried them to San Francisco. As in the rest of the country, the railroad's importance as a passenger carrier declined after World War II.

Cable cars

The instantly recognizable symbol of San Francisco is the cable car. Invented by Andrew Hallidie (because he felt sorry for the dray horses that were often injured on the steep hills), the system was tested in 1873. Other cities eventually abandoned cable cars, but San Francisco has stubbornly clung to the picturesque if archaic, and sometimes dangerous, means of negotiating the hills. Before the 1906 earthquake 600 cars covered 110 miles (177 km) of the city, but the system was devastated by the quake and much of it was not restored. Today more than two dozen cars operate at peak hours, carrying about 25,000 people daily to limited destinations via three lines.

San Francisco International Airport is located about 7 miles (11 km) south of the city-county limits, occupying a filled site on the southwestern shore of the bay.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. San Francisco is coextensive with San Francisco county. Unlike any other California city, it (incorporated 1850) has a consolidated city-county government. The 1932 freeholders' charter, under which the city-county still operates, provides the mayor with strong executive powers but delegates substantial authority to a chief administrative officer (appointed by the mayor) and a controller. The legislative authority is lodged with an

elected board of supervisors. The other key officials, who are both appointed, are the superintendent of schools and the manager of utilities.

Public utilities. Since 1934 San Francisco's principal source of water has been the Hetch Hetchy Reservoir, 167 miles (269 km) away, in the Sierra Nevada. Other sources are the Calaveras Reservoir in Alameda county and reservoirs in San Mateo county to the south. The Hetch Hetchy project required the damming of a scenic valley in Yosemite National Park and the construction of tunnels, one 25 miles (40 km) long, through the Coast Range. In 1902 the first high-voltage line transmitting hydroelectric power was completed between a powerhouse on the Mokelumne River and San Francisco, some 180 miles (290 km) in length. Since then, the Bay Area has developed a network of hydroelectric plants on the rivers of the interior, as well as a steam-powered plant on Monterey Bay.

Education. The Bay Area is one of the country's centres of higher learning. Although strictly speaking they cannot be counted as San Francisco institutions, two of the region's universities—the University of California, located across the bay in Berkeley (campus opened 1873), and Stanford University (opened 1891), neighbour to Palo Alto down the peninsula—are among the nation's most prestigious schools. Within San Francisco itself are the University of San Francisco, originally a Jesuit academy established in 1855, and San Francisco State University, which was founded as a normal school in 1899, became a four-year college in 1935, and achieved university status in 1972. Other institutions include Golden Gate University (1853), the City College of San Francisco (1935; a two-year public college), and the San Francisco Art Institute (1871).

CULTURAL LIFE

The arts. A great part of San Francisco's appeal is its well-established image as a cultural centre. By 1880 it boasted one of the largest opera houses in the country, the largest hotel, a public park, great churches and synagogues, and a skyline bristling with the mansions of millionaires. Drama and music flourished there, with appearances by such luminaries as Sarah Bernhardt, Edwin Booth, Luisa Tetrazzini, James O'Neill, Lillie Langtry, and Lotta Crabtree. San Francisco native Isadora Duncan began her career there teaching modern dance.

Literary San Francisco

The city's true artistic calling, however, has been as a mecca for writers. One of the first was Mark Twain, who arrived in time for the great silver boom that came some 10 years after the gold boom faded. Other noted writers were Ambrose Bierce, who came to the city after horrendous experiences in the American Civil War, Jack London, Bret Harte, Frank Norris, Gertrude Atherton, and Robert Louis Stevenson, who lived in great poverty in a boardinghouse; later came Dashiell Hammett, Stewart Edward White, Kathleen Norris, Erskine Caldwell, William Saroyan, and Wallace Stegner. During the mid-1950s, San Francisco became known as a centre of the Beat movement, and poet Lawrence Ferlinghetti's City Lights Bookstore, which was the country's first to sell paperbacks, became one of the movement's best-known gathering places. More recent Bay Area authors are Amy Tan, Herbert Gold, Anne Lamott, Ethan Canin, and Danielle Steele.

San Francisco is home to two major musical institutions. The San Francisco Symphony performs in the Louise M. Davies Symphony Hall and gives pop concerts in the summer. The San Francisco Opera stages an early season to allow its leading singers to fulfill their commitments at New York City's Metropolitan Opera. With the exception of the American Conservatory Theater (A.C.T.), a resident repertory group, the professional theatre is virtually nonexistent in the city. The surviving downtown theatres are largely occupied by the touring casts of Broadway shows.

San Franciscans believe their city is a haven for the artist. While this would hold true for those who value architecture and public sculpture, the painting collections do not rival those of Los Angeles or the East Coast. Notable, however, are the jades and porcelains in the Asian Museum, the Rodin sculptures at the California Palace of the Legion of Honor, the downtown Museum of Modern Art, and the many treasures in such small museums as the Fire Department Pioneer Memorial Museum. While San Francisco's artistic community does not approach the prominence of its writing establishment, it has produced such notable figures as Wayne Thiebaud and Richard Diebenkorn.

Cultural institutions. Several cultural institutions were constructed after the 1906 earthquake, among them the Civic Center (a lovely square sparkling with fountains surrounded by such Renaissance revival-style buildings as City Hall), the public library, and the civic auditorium. Publisher M.H. de Young helped fund the building of the M.H. de Young Museum in Golden Gate Park, and Adolph and Alma de Bretteville Spreckels sponsored the stately California Palace of the Legion of Honor, which overlooks the Golden Gate Bridge. A spectacular reminder of the 1915 Panama-Pacific International Exposition is found in the monumental Palace of Fine Arts, located in a little park in the Marina District. Housing the Exploratorium (a science museum), the palace is a giant Neoclassical rotunda, designed by the architect Bernard Maybeck and completely restored in the 1960s.

Museums

Popular culture. A vital part of San Francisco culture is found in its restaurants, bars, hotels, and clubs. To this must be added the popular culture of the ethnic enclaves.

In the minds of many, however, San Francisco's most memorable contribution to the nation's culture is its past. In the late 1960s the Haight-Ashbury District became a haven for the "flower children" and "hippies" who declared themselves in headlong flight from established society and who preached the saving graces of peace, love, and hallucinogens. However, by the 1970s Haight-Ashbury had become an ugly and dangerous marketplace for drugs and vice. Though an occasional homeless person, drugged beggar, or aging hippie can still be encountered, with the rise in real estate prices all over the city, Haight-Ashbury now boasts a middle-class population and specialty boutiques, upscale restaurants, used bookstores, and the ubiquitous coffeehouses.

History

EXPLORATION AND EARLY SETTLEMENT

It is extraordinary that the site of San Francisco should have been explored first by land instead of from the sea, for San Francisco Bay is one of the most splendid natural harbours of the world, yet great captains and explorers—Juan Rodríguez Cabrillo (1542–43), Sir Francis Drake (1579), and Sebastián Vizcaíno (1602)—sailed unheeding past the entrance. In 1769 a scouting party from an expedition led by the Spanish explorer Gaspar de Portolá looked down from a hilltop onto a broad body of water; they were the first Europeans known to have seen San Francisco Bay. It was not until Aug. 5, 1775, that the first Spanish ship, the *San Carlos,* commanded by Lieut. Juan Manuel de Ayala, turned eastward between the headlands, breasted the ebbing tide, and dropped anchor just inside the harbour mouth. It is possible that Drake may have entered the bay, but most evidence suggests otherwise.

Settlers from Monterey, under Lieut. José Joaquin Moraga and the Rev. Francisco Palóu, established themselves at the tip of the San Francisco peninsula the following year. The military post, which remained in service as the Presidio of San Francisco until 1994, was founded in September 1776, and the Mission San Francisco de Asis, popularly called the Mission Dolores, was opened in October.

Mission Dolores

Almost half a century later, a village sprang up on the shore of Yerba Buena Cove, 2 miles (3 km) east of the mission. The pioneer settler was an Englishman, Capt. William Anthony Richardson, who in 1835 cleared a plot of land and erected San Francisco's first dwelling—a tent made of a ship's foresail. In the same year, the United States tried unsuccessfully to buy San Francisco Bay from the Mexican government, having heard reports from whalers and captains in the hide-and-tallow trade that the great harbour held bright commercial possibilities.

The Americans had to wait only 11 years. After fighting began along the Rio Grande, Capt. John B. Montgomery sailed the sloop of war *Portsmouth* into the bay on June 3, 1846, anchored in Yerba Buena Cove, and went ashore with a party of sailors and marines to raise the U.S. flag in

the plaza. On Jan. 30, 1847, Yerba Buena was renamed San Francisco, which was regarded as more propitious.

The permanent European population of Yerba Buena in 1844 did not exceed 50 persons. By 1846 the settlement had a population of 375, in addition to 83 African Americans, Native Americans, and Sandwich Islanders (Hawaiians). Two years later, just before the discovery of gold on the American River, the town had grown to about 200 shacks and adobes inhabited by about 800 settlers.

THE GROWTH OF THE METROPOLIS

City of the '49ers

With the discovery of gold at Sutter's Mill in the Sierra Nevada, San Francisco picked up pace and direction. The modest village was at first almost deserted as its population scrambled inland to the Mother Lode, and then it exploded into one of the most extraordinary cities ever constructed. Some 40,000 gold hunters arrived by sea, 30,000 plodded across the Great Basin, and another 9,000 moved north from Mexico. By 1851 more than 800 ships rode at anchor, deserted by their crews.

Everybody except the miners got rich. Eggs sold for a dollar apiece, and downtown real estate claimed prices that would almost hold their own against the current appreciated values. Until the bubble burst in the panic of 1857, 50,000 San Franciscans became rich and went bankrupt, cheated and swindled one another, and took to violence all too readily. As *The Sacramento Union* noted in 1856, there had been "some fourteen hundred murders in San Francisco in six years, and only three of the murderers hung, and one of these was a friendless Mexican." Two vigilance committees responded with crude and extralegal justice, hanging eight men as an example to the others.

In 1859 silver was discovered in the Nevada Territory. The exploitation in Nevada of the Comstock Lode, which eventually yielded some $300 million, turned San Francisco from a frontier boomtown into a metropolis whose leading citizens were bankers, speculators, and lawyers, all of whom ate and drank in splendid restaurants and great hotels. By 1870 San Francisco boasted a population of nearly 150,000. The 1860s and '70s marked the birth of the modern San Francisco, which has since then claimed to be the Athens, Paris, and New York City of the West but has never completely lost its mark of a wild beginning.

THE 20TH CENTURY

Expansion during the world wars. While the rest of the world was preparing for World War I, San Francisco held a highly successful World's Fair—the Panama-Pacific International Exposition—to celebrate the new boost to Western commerce, the opening of the Panama Canal. During the Great Depression, 4,000 longshoremen competed for 1,300 jobs parceled out by the International Longshoremen's and Warehousemen's Union. The ILWU fought scabs and union busters at the port on "Bloody Thursday," July 4, 1934, and then called a citywide general strike, the largest in the country's history.

World War II made a significant impact on San Francisco's prosperity, as it served as a major transfer point for the Pacific theatre. Great shipyards were built around the bay, and some half million people came to work in the area's war-related industries; many of them stayed on permanently after the war. The United Nations was born there in 1945, the result of the San Francisco Conference, which took place that year from April to June.

From peace to protest. San Francisco in the 1950s was remarkable, not only for its role in the Beat movement but for the number of performers who came to fame in its clubs and cafés: Lenny Bruce, Jonathan Winters, Woody Allen, Phyllis Diller, Barbra Streisand, and Mort Sahl all had their first successes in North Beach venues. The next decade was marked by drugs, hippies, and violent protests against the Vietnam War. As one wag has said, "If you can remember the '60's in San Francisco, you weren't there." The city emerged as a centre of psychedelic rock music, with such local groups as the Jefferson Airplane, Grateful Dead, and Quicksilver Messenger Service, as well as such individual performers as Janis Joplin. The city also at that time became a centre for environmentalists and advocates of gay and minority rights. San Francisco was one of the first cities in the country to bus students in order to achieve racial integration; the Save the Bay Association and San Francisco Bay Conservation and Development Commission were formed in the mid-1960s; and in 1969 a group of Native Americans, believing they had a right to unused government land, invaded Alcatraz Island and occupied it until 1971.

The 1960s

Several violent acts put the city in the news in the 1970s. In September 1975 an assassination attempt was made against President Gerald Ford in a downtown square, and in November 1978 the followers of Jim Jones (whose cult-like ministry was based in San Francisco) died in a mass suicide in Jonestown, Guyana. A few days after the Jonestown Massacre, Mayor George Moscone and City Supervisor Harvey Milk were murdered at City Hall. These events had a sobering effect on the city, in contrast to the freewheeling atmosphere of the previous decade. However, Dianne Feinstein, the city's first female mayor, provided crucial stability after Moscone's assassination. San Francisco completed BART, its rapid transit system, in the 1970s and established the Golden Gate National Recreation Area, some 110 square miles (285 square km) in San Francisco, Marin, and San Mateo counties.

The late 20th century. San Francisco experienced great growth in the 1980s. The city's population topped 700,000, not least because of the great influx of immigrants from South Asia. The cost of living skyrocketed, making San Francisco one of the most expensive cities in the country. The number of automobiles doubled, the popular but deteriorating cable cars received a multimillion-dollar facelift, tourism became the city's most lucrative business, and the city's homeless population grew precipitously, as it did throughout the United States. But by far the most momentous event locally, if not nationally, was the earthquake of 1989. A milestone was reached in 1995 when the city's first African American mayor, Willie L. Brown, Jr., was elected.

1990s prosperity

As the century came to a close, the city continued to face a multitude of urban problems, from affordable housing, crime, and homelessness to pollution, traffic, and the assimilation of new immigrants. The homosexual community continued to struggle against what it perceived to be an inadequate if not indifferent response from the government to the AIDS crisis. In 1997 San Franciscans held a candlelit vigil following the death of Pulitzer Prize-winning columnist Herb Caen. The "cool grey city of love" had been Caen's bailiwick for more than 60 years, and with his death San Francisco lost one of its favourite sons.

(K.La./G.C.Ha./B.Co./Ed.)

BIBLIOGRAPHY. For accounts of the early history of San Francisco, FRANK SOULÉ, JOHN H. GIHON, and JAMES NISBET, *The Annals of San Francisco* (1855), is invaluable; whereas B.E. LLOYD, *Lights and Shades in San Francisco* (1876), is both vivid and divertingly moralistic. HERBERT ASBURY, *The Barbary Coast* (1933, reprinted 1968), is a classic account of the underworld; and JULIA COOLEY ALTROCCHI, *The Spectacular San Franciscans* (1949), is a useful social history. JOHN HASKELL KEMBLE, *San Francisco Bay: A Pictorial Maritime History* (1957, reprinted 1978), contains splendid drawings and photographs. WILLIAM BRONSON, *The Earth Shook, the Sky Burned* (1959, reprinted 1997), is a first-rate historical account of the 1906 earthquake.

The growth of San Francisco is treated in GUNTHER BARTH, *Instant Cities: Urbanization and the Rise of San Francisco and Denver* (1975, reissued 1988). FREDERICK M. WIRT, *Power in the City: Decision Making in San Francisco* (1974, reissued 1978), is a study of local politics. The WRITERS' PROGRAM, CALIFORNIA, *San Francisco: The Bay and Its Cities*, new rev. ed. (1973), is a well-known guide; HAROLD GILLIAM, *San Francisco Bay* (1957), written by the naturalist-conservationist, is authoritative and evocative. Comprehensive books on the city's history include GLADYS HANSEN, *San Francisco Almanac: Everything You Want to Know About the City*, updated and rev. ed. (1980); and LAWRENCE FERLINGHETTI and NANCY J. PETERS, *Literary San Francisco: A Pictorial History from Its Beginnings to the Present Day* (1980). MORTON BEEBE, *San Francisco*, new rev. ed. (1993, reissued 1996), includes essays by Herb Caen, Tom Cole, Barnaby Conrad, Herbert Gold, Kevin Starr, and John Hart, as well as vivid photographs of the city. BARNABY CONRAD (ed.), *The World of Herb Caen* (1997), provides a lively account of the life of the newspaper columnist from 1938 to 1997. RICHARD SAUL WURMAN and DONNA PECK, *Access San Francisco*, 8th ed. (1999), is a helpful guidebook. (B.Co.)

São Paulo

The largest city of Brazil and dynamic capital of the state of the same name, São Paulo is the foremost industrial centre in Latin America. With one of the world's fastest growing metropolitan populations, it is also the largest city of the Southern Hemisphere and one of the largest conurbations in the world. Sometimes called the "locomotive that pulls the rest of Brazil," São Paulo has a vibrant and energetic urban core characterized by an ever-growing maze of modern steel, concrete, and glass skyscrapers. The city is located in the hills of the Serra do Mar, which forms part of the Great Escarpment that extends between the Brazilian Highlands and the Atlantic Ocean. It lies about 220 miles (354 kilometres) southwest of Rio de Janeiro and about 30 miles inland from the port of Santos. The city's name derives from its having been founded by Jesuit missionaries on January 25, 1554, the anniversary of the conversion of St. Paul.

The article is divided into the following sections:

Physical and human geography 6
The landscape 6
The city site
Climate
The city plan
The people 6
The economy 7
Industry and commerce
Transportation
Administration and social conditions 8
Government and services
Education
Cultural life 8
History 8
Bibliography 9

Physical and human geography

THE LANDSCAPE

The city site. The Brazilian Highlands are composed of ancient crystalline rocks, which in the vicinity of São Paulo form a surface of gently rounded hills mantled with a reddish clay soil. Rivers such as the Tietê, on which São Paulo is located, rise near the edge of the Great Escarpment and flow generally westward to the Rio Paraná. In their course, they cross stratified sandstones and limestones overlaying the crystalline base, as well as sheets of volcanic rock that form the Paraná Plateau. Here, there are rapids and waterfalls, as well as dams and reservoirs that supply great quantities of hydroelectric power.

Located at an elevation of 2,690 feet (820 metres) above sea level, the city is surrounded by open country, valleys, and foothills. The higher terrain constitutes the preferred residential areas; the lower parts are on alluvial land along the banks of three rivers (the Tietê, the Pinheiros, and the Tamanduateí), and these are occupied by working-class residences, manufacturing establishments, and commercial enterprises. The area of the city is 576 square miles (1,493 square kilometres), but including suburban communities, such as Santo André, Diadema, São Bernardo do Campo, São Caetano do Sul, Osasco, Guarulhos, Mairiporã, Barueri, Santana do Parnaíba, Franco da Rocha, and Mogi das Cruzes, metropolitan São Paulo sprawls over an area of 3,070 square miles. Open spaces on the perimeter of the city, where there are clay soils mixed with sandy deposits, are used for intensive market gardening. A forest reserve of about 39 square miles is maintained in the nearby Serra da Cantareira, while the beaches of Santos and Guarujá provide pleasant resort areas.

Climate. The Tropic of Capricorn, at about 23°27′ S, passes through São Paulo and roughly marks the boundary between the tropical and temperate areas of South America. Because of its elevation, however, São Paulo enjoys a distinctly temperate climate. July is the coldest month, with an average temperature of 57.9° F (14.4° C) and occasional frost. Warmest is February, which averages 69.1° F (20.6° C). Rainfall is abundant, particularly during the summer season from October through March, averaging 56 inches (1,422 millimetres) per year. Humidity and air pollution combine to form a mist that often hangs over the city.

The city plan. The central business district of São Paulo focusses on the famous Triângulo, where the 42-story Edifício Itália, inaugurated in 1975, rises to a height of 558 feet. In 1947 there were only three skyscrapers in all of São Paulo, among which the newly constructed, 36-story Banco do Estado de São Paulo was the tallest. Now, the entire city is studded with modern buildings whose construction reflects a variety of architectural styles and materials.

Surrounding the central business district are extensive areas devoted to manufacturing, wholesale and retail trade, and repair and maintenance services. Most extensive are the residential areas characterized by low, red-roofed houses, interspersed with high-rise apartments or office complexes, singly or in clusters. Suburban shopping centres, like suburban neighbourhoods, have become commonplace.

São Paulo had no city plan until 1889, and no zoning law was passed until 1972. Until well into the 19th century, therefore, this state capital retained a colonial aspect, with narrow, unpaved streets, shabby buildings, and old churches and convents of Jesuit and Franciscan styles. Successive city administrations since then have attempted to stimulate more rational urban growth and to modernize the city's transportation system. Projects have included the straightening and rechannelling of the Tietê and Tamanduateí rivers, the widening and relocation of streets and avenues, the development of new parks and lakes, and the construction of superhighways and of a 40-mile subway network. Still, the city's street pattern shows little overall coherence, and the problems of intraurban traffic congestion and pollution have reached monumental proportions.

THE PEOPLE

The original settlers of São Paulo were relatively poor and largely from southern Portugal. They were, however, a restless people who sought actively to improve their status in life. Among them were the *bandeirantes* ("explorers") who formed expeditions that pushed far into the interior of South America in search of slaves and mineral wealth and, in the process, expanded the frontiers of what has become modern-day Brazil.

With the great expansion of coffee cultivation in São Paulo state after 1880 came a massive immigration of Europeans. Italians and Portuguese were the most numerous, but there were also many Spaniards, Germans, and eastern Europeans. Other settlers came from Japan and the Middle East. Today, there are more Japanese in São Paulo than in any other community outside Japan, and Japanese farmers supply much of the city's market for fruit and vegetables. Even more numerous are internal migrants, primarily from the Northeast of Brazil. These include many blacks, who are the descendants of African slaves. Overall, the population is more than half European and about one-third black and mulatto, with the remainder made up of small groups of Asians and others. Roman Catholicism is the near-universal religion, and the archdiocese of São Paulo is the world's largest in number of adherents. Various other religions are represented by

São Paulo metropolitan area.

smaller numbers, and many Paulistas, as the inhabitants of the city are known, also attend the rites of local cults. Portuguese is the predominant language, although other languages are commonly spoken.

THE ECONOMY

Industry and commerce. Industrial development, beginning in the late 19th century, but especially since World War II, has transformed metropolitan São Paulo into the foremost industrial centre in Latin America. This city has often been referred to as the "Chicago of South America," but it actually is a leader of Brazilian commerce and industry to a much greater degree than is Chicago in the United States. The value of its industrial production is by far the country's largest. Its leading industries produce textiles, mechanical and electrical appliances, furniture, foodstuffs, and chemical and pharmaceutical products. There are also heavy metallurgical plants located at nearby Taubaté, oil refineries and chemical plants at Cubatão, and plants manufacturing motor vehicles and farm machinery in São Bernardo do Campo, Santo André, and other suburban communities. The several thousand manufacturing establishments in São Paulo provide employment for some 15 percent of the population. Despite its rapid growth in recent decades, however, the industrial sector has been able to absorb only a small fraction of the growing labour force. Hence, unemployment and underemployment are continuing problems.

Commerce, both wholesale and retail, is well developed and is spread over the city by zones according to specialty. Banks are concentrated in the central Triângulo of the city but maintain branches in almost every district. In addition to important Brazilian banks, there are banking institutions representing interests in North and South America, Europe, Asia, and Africa. No less important, in terms of employment, are street vending, peddling, and neighbourhood stores.

Transportation. Major arteries of transportation radiate in all directions from São Paulo. Three major airports—Congonhas within the city, Cumbicas 15 miles east, and Viracopos 60 miles northwest—together with several smaller ones, provide São Paulo with both international and domestic service. The Viaçao Aérea São Paulo (VASP), with headquarters in the city, is Brazil's second largest airline and is owned by São Paulo state. Marine transport is provided through the port of Santos. São Paulo is also a hub of railroads, which include a transcontinental line from Santos to Antofagasta, Chile. Modern highways connect with inland cities, Santos, Rio de Janeiro, and almost all the states of Brazil. Within the city, the first freeway was opened in 1969, and the subway system was inaugurated in 1976. Automobile traffic in the city and

Museu Paulista da Universidade de São Paulo (Paulista Museum) in São Paulo.
Colour Library International

suburbs is heavy, and, despite street and highway improvements, congestion is a major and growing problem, which adds to the industrial city's serious conditions of air and noise pollution.

ADMINISTRATION AND SOCIAL CONDITIONS

Government and services. São Paulo is governed by a mayor and city council. It is also the seat of state government, headquartered in the Palácio dos Bandeirantes in the southwestern district (neighbourhood) of Morumbi. Many state offices and departments are headquartered in the city, as are many branches of the federal government. Dozens of countries maintain consulates there.

The city and state have constructed a chain of various reservoirs, tunnels, and canals to supply fresh water to a metropolitan population of nearly 20,000,000 people. The Cantareira water supply project, begun in 1969, has increased water supplies greatly, but demand has continuously outstripped supply and made it necessary to undertake more large-scale projects. Pollution has been an ever-present danger because dammed streams that carry industrial waste run slower, though efforts have been made to clean up the Rio Tietê.

Hydroelectric power

Electricity has been available in abundance to São Paulo since 1900. First, the waters of the Rio Tietê were dammed and dropped through penstocks from the Great Escarpment to generators below. Subsequently, dams on many rivers to the west, including Itaipú, a joint project with Paraguay and one of the largest hydroelectric dams in the world, have been built to sustain the city.

Public and private health facilities are numerous, including hospitals for civil servants, maternity hospitals, and hospitals specializing in the treatment of cancer, heart diseases, tuberculosis, and other illnesses.

Education. São Paulo has a well-developed system of primary and secondary education, both public and private, and a variety of vocational-technical schools. Among the institutions of higher education, the largest and most esteemed in all of Brazil is the state-supported Universidade de São Paulo, established in 1934, which incorporated the historic Faculdade de Direito (College of Law) in the old São Francisco Square and preexisting polytechnical schools, as well as schools of pharmacy, dentistry, agriculture, and medicine. Economics, architecture, and engineering were added later. Affiliated institutions include a school of sociology and politics, founded in 1933, and the Instituto Butantã, a world-famous centre for research on snakes and the production of antitoxins and antivenins. The Pontifícia Universidade Católica de São Paulo was established in 1946, and the Universidade Mackenzie in 1952. Also well known is the Escola de Administração de Emprêsas of the Getúlio Vargas Foundation, which offers business administration and economics programs.

CULTURAL LIFE

São Paulo became a prominent cultural and intellectual centre in the 19th century, largely due to the opening in 1827 of the Faculdade de Direito, one of the first two in Brazil, where many eminent leaders were educated. The Instituto Histórico e Geográfico de São Paulo, founded in 1894, is one of the oldest cultural associations in the state. The city is also a leading centre for libraries, publishing houses, and theatres. The municipal library is housed in one of São Paulo's skyscrapers. In 1922 São Paulo's Modern Art Week, celebrated by a group of young writers, artists, and musicians in the Teatro Municipal, introduced modernism in the arts of Brazil. The Museu de Arte de São Paulo, founded in 1947, is one of the best in South America, and the Museu de Arte Contemporânea is also outstanding. São Paulo's symphony orchestra is similarly advanced in the field of music.

Museums and libraries

Publishing and broadcasting have long been established in São Paulo. Several of the nation's largest and most influential newspapers are published in the city, including *Fôlha de São Paulo* (1921) and *O Estado de São Paulo* (1875). Television was introduced in 1950, and the city is headquarters for some of the most important Latin-American radio stations.

The Paulistas are noted for their enthusiasm for sports. Football (soccer) is the predominant sports attraction, as evidenced by the large-capacity Morumbi and Pacaembu stadiums. Also popular are swimming, tennis, volleyball, basketball, and auto racing, for which São Paulo has one of the world's largest tracks, at Interlagos on the city's south side. There are also countless parks, plazas, and playgrounds. The São Paulo zoo (1958), with some 3,200 animals, is one of the world's largest.

History

São Paulo was the first highland settlement established in Brazil. It began as a small Indian settlement in 1554 under the direction of Portuguese Jesuit missionaries and occupied the lower terraces of the Rio Tietê in the midst of tall grasses and scattered scrub trees. The community grew slowly and had only 300 inhabitants by the end of the 16th century. Yet, São Paulo became a township in 1560 and had a town council that could enact and enforce laws. In 1683 it succeeded São Vicente as seat of the captaincy, or hereditary fief, and the inhabitants already had become known as Paulistanos or Paulistas.

Early settlement

Seventeenth-century São Paulo was a base for expeditions (*bandeiras*) into the hinterlands of armed pioneers (*bandeirantes*) in search of Indian slaves, gold, silver, and diamonds. In the process they expanded the frontiers of what was to become modern-day Brazil. In 1711 São Paulo attained the status of a city, yet it remained an agrarian town that had yet to see any significant prosperity.

THE CITY AFTER INDEPENDENCE

The Portuguese regent Dom Pedro (later Pedro I) declared Brazil's independence on September 7, 1822, on the plain of Ipiranga, which is now within São Paulo. By 1840 São Paulo was still a town of 20,000 inhabitants centred on a low hill and its neighbouring Anhangabaú valley. The Tamanduateí River was straightened in 1849 and a municipal market constructed in 1867. But only in 1875 was the colonial centre linked by a new street to the square now called the Praça da República. By then brick houses were being built, and gas streetlamps and horse-drawn streetcars were coming into use. (A.Le./C.W.M.)

As coffee became Brazil's main source of export earnings, São Paulo and Santos, its port, grew at a spectacular rate. Italians, who accounted for more than 600,000 of the nearly 900,000 foreigners coming to the state between 1888 and 1900, soon came to outnumber native Brazilians. The ethnic mosaic was further enriched by Portuguese, Spaniards, Germans, and eastern Europeans, followed by Syrians, Lebanese, and Japanese. Coffee planters' townhouses sprang up in Higienópolis and Campos Elíseos in the west, and crowded workers' housing extended through Mooca and Brás in the east. Three-story buildings began to appear on the central Triângulo ("Triangle"), and in 1892 an iron viaduct was built across the Anhangabaú valley from Cha Hill, where tea (*cha*) plants had been grown only a few years earlier. The population jumped from 44,000 in 1886 to nearly 130,000 by 1893.

New industries by 1905 included textile mills, shoe factories, and others using local raw materials. Cotton textile mills alone employed 39,000 workers. During the period 1899–1911 Mayor Antônio Prado widened streets, completed the monumental Luz train station, and started construction of the Santa Ifigênia Viaduct, opened in 1913. His successor remodeled the Praça da Sé, created the Praça da Patriarca, and completed Dom Pedro II Park, begun in 1911. By that time the opulent mansions of the wealthiest coffee barons lined Paulista Avenue, and concrete buildings of five to six stories were becoming common in the city centre, which was crowned by a seven-story marvel of reinforced concrete. São Paulo's population grew from about 240,000 in 1900 to 580,000 by 1920.

FROM CITY TO METROPOLIS

Physical and demographic changes. São Paulo maintained a high growth rate through the 1920s, driven by interrelated streams of immigration, rapid industrialization, and investment. In the early 1920s the Sampaio Moreira Building reached an unprecedented 14 stories, and by the end of the decade the Martinelli Building attained nearly twice that height. Growing fleets of automobiles and diesel buses ferried hordes of commuters from their outlying homes to jobs in the city centre. Although a modern face had come to São Paulo's better areas by the 1930s, larger portions were basically unchanged. Indeed, well into the 20th century much of the city retained a colonial aspect, with narrow, unpaved streets, shabby buildings, and a few old churches of Jesuit and Franciscan styles.

Between 1920 and 1940 the population more than doubled, reaching 1.3 million. During 1939–45 the engineer-mayor Francisco Prestes Maia built 9 de Julho ("9th of July") Avenue and widened numerous other streets despite resistance from displaced residents. By 1947 the new star of São Paulo's skyline was the 36-story São Paulo State Bank building, and, starting with the Mário de Andrade Municipal Library, the city's architecture moved beyond the short period of Art Deco design.

By 1950 São Paulo had grown to a metropolis of 2.2 million souls compared to Rio's 2.4 million, but a decade later São Paulo led by 3.7 million to Rio's 3.3 million, thus solidifying its reputation as one of the world's most dynamic urban centres. Famed architect Oscar Niemeyer was lured from Rio to design the sinuous curves of the Copan Building (1952), and the 42-story Itália Building (1956) became its towering neighbour. The highly imaginative São Paulo Art Museum (1960) was built over the juncture of 9 de Julho and eight-lane Paulista Avenue.

In the 1960s São Paulo came to include almost half of the population of São Paulo state (Brazil's most populous state) and to account for about one-third of the county's total industrial employment. By 1970 São Paulo city had leapt to nearly 6 million inhabitants, and more than 2 million more lived in its suburbs. Because automobiles were becoming a São Paulo family staple, expressways were built along the canalized Tietê and Pinheiros rivers in 1967. However, no amount of highway construction and street widening could more than briefly alleviate traffic congestion. A subway system was constructed (1965–79) in hopes of improving the situation.

The political mix. In the latter part of the 20th century São Paulo governors and mayors, seeking to springboard themselves to national office, began to emulate Mayor Prestes Maia by undertaking sorely needed public works programs. The bon vivant Adhemar de Barros, who was the state's appointed chief executive during 1938–41, subsequently won elections for mayor and governor on the basis of such projects as the Anchieta and Anhanguera expressways, the massive Hospital das Clínicas, the electrification of the Sorocabana railroad, and the Vila Leopoldina sewage treatment plant. These accomplishments earned him the slogan "He may rob, but he gets things done" and helped his presidential campaign in 1955. However, de Barros was defeated by Juscelino Kubitschek, who had achieved as much, but with no taint of graft, as mayor of Belo Horizonte and governor of Minas Gerais.

The populist reformer Jânio da Silva Quadros served as mayor in 1953–54 and governor in 1955–58. He finished Prestes Maia's street-building plans, straightened the Tietê and Pinheiros rivers, and built hydroelectric projects to meet the city's insatiable appetite for power. De Barros, who subsequently reclaimed the mayor's office (1957–61) and the governorship (1962–66), completed some projects begun by his rival and initiated construction of the subway and the Tietê and Pinheiros expressways.

FROM METROPOLIS TO MEGAMETROPOLIS

Under the military regime of 1964–85 São Paulo's economy grew, and additional public works projects were initiated. However, *favelas* also began to envelop parts of the city during the 1960s and '70s, when as many as 300,000 people—many of them from Brazil's impoverished Northeast—poured into the metropolitan region each year. The conservative Paulo Salim Maluf, who served both as mayor (1969–71) and governor (1979–82), extended water and sewer services, removed *favelas* from central areas, and built public housing complexes on the periphery.

Quadros, who won the 1985 mayoral elections, promoted public safety and leveled *favelas* to make way for new construction. In 1989–93 the radical reformer Luíza Erundina concentrated on social welfare and low-cost housing, notably in the *favelas*, but Maluf, who was reelected in 1993, ended those policies. Following Celso Pitta's mayoral term (1997–2001), which was marred by corruption scandals, Marta Suplicy occupied the office.

In the early 21st century São Paulo retained its status as one of the world's most populous urban areas, with some 10 million people living in the city proper and additional millions in the suburbs. However, more than one million resided in *favelas*, traffic congestion plagued the streets, public services fell short of demand, and the rates of murder and other violent crimes were disproportionately high for Brazil. Nonetheless São Paulo remained a dynamic and vital component of Brazilian national life.

BIBLIOGRAPHY. Studies of geography, population, economics, and politics include NICE LECOCQ MÜLLER, "Demographic Growth and Urban Expansion in the Metropolitan Area of São Paulo," in *Revista Geográfica*, no. 97, pp. 29–30 (January–June 1983); and JOSEPH L. LOVE, *São Paulo in the Brazilian Federation: 1889–1937* (1980).

RICHARD M. MORSE, *From Community to Metropolis: A Biography of São Paulo, Brazil*, new and enlarged ed. (1974), is a standard history. Analyses of ethnic relations, poverty, and social problems include GEORGE REID ANDREWS, *Blacks & Whites in São Paulo, Brazil, 1888–1988* (1991); MICHAEL GEORGE HANCHARD, *Orpheus and Power: The Movimento Negro of Rio de Janeiro and São Paulo, Brazil, 1945–1988* (1994, reissued 1998); and TERESA PIRES CALDEIRA, *City of Walls: Crime, Segregation, and Citizenship in São Paulo* (2000), which focuses on violence in the city. (Ed.)

Scandinavian Literature

Scandinavian literature consists of those writings in the North Germanic group of the Germanic languages, the modern forms of which include Swedish, Norwegian, Icelandic, Danish, and Faeroese. The literary works written in these languages, though manifesting certain differences reflective of distinct national institutions, exhibit strong similarities stemming from deep-seated common linguistic and cultural ties. Some authorities include Finland among the Scandinavian countries on geographical and economic grounds, but the literature of the Finnish-speaking people, like their language, stands apart in a number of respects. (Finnish belongs to the Baltic-Finnic branch of the Finno-Ugric language family and is most closely related to Estonian, Livonian, Votic, and Karelian.) The present article does, however, devote some attention to various notable Finnish authors who wrote in Swedish. (Ed.)

The article is divided into the following sections:

The Middle Ages 10
- Norwegian and Icelandic literature 10
  - The classical period in Iceland
  - Post-classical literature in Iceland
- Swedish literature 12
- Danish literature 12

The 16th century 13
- The impact of the Reformation on Swedish letters 13
- Developments in Danish literature 13
- Icelandic learning and literature 13

The 17th century 13
- Swedish poetry and prose 13
- The literary Renaissance in Denmark 13
- Renewed literary activity in Norway 14
- Icelandic letters 14

The 18th century 14
- Swedish Classicism and Enlightenment 14
- Literary activity in Denmark, Norway, and Iceland 14

The 19th century 15
- Swedish literature 15
  - Romanticism
  - Emergence of Realism and Poetic Realism
  - Sources of modern Swedish literature
  - Finno-Swedish literature
- Norwegian literature 16
  - The Age of Wergeland
  - National Romanticism
- Danish literature 16
  - The Romantic period
  - Romantic Realism
  - Poetic Realism
- Faeroese literature 17
- Icelandic literature 17

The 20th century 17
- Norwegian 17
- Swedish 18
  - The modern Swedish novel
  - Development of lyric poetry
  - Contemporary trends
  - Developments in Finno-Swedish literature
- Danish 19
  - The influence of Georg Brandes
  - Neoromantic revival
  - 20th-century literary trends
- Faeroese 20
- Icelandic 20
  - Notable 20th-century prose writers
  - Major poets
  - The development of the Icelandic drama

Bibliography 21

The Middle Ages

The literature of Scandinavia and, in particular, of Iceland has reflected two extraordinary features of the social and cultural history of pagan Europe and of Iceland. The way in which names such as Siegfried, Brunhild, and Attila cropped up again and again in different European literatures has borne witness to the dissemination of legends and traditions common to the early Germanic tribes of Europe, starting from the great movements westward in the 4th, 5th, and 6th centuries. The literature of Iceland provides not only the most detailed descriptions available of the life-style of early Germanic peoples but constitutes the most complete account of their literature and literary traditions. Although the sagas and poems were first written down by Christian scribes, they present a picture of a pre-Christian European culture that reached its heights in the new settlements in Iceland.

A second feature directly concerns the peoples of Scandinavia. A remarkable characteristic of Scandinavian literature was the accuracy with which it described the geography of northern Europe, accuracy that was born of actual knowledge. From the late 8th century until well into the Middle Ages, the history of the Norsemen was one of unceasing movement toward western and central Europe. The Norsemen discovered Iceland, as early Icelandic historians had it, when their ships were blown off course about 860. The next century found the Vikings pushing west by way of Britain, Ireland, and France to Spain and then through the Mediterranean to North Africa and east to Arabia. Across land they reached the Black Sea, by sailing north they came to the White Sea, and finally, turning westward again, they reached America long before Columbus.

NORWEGIAN AND ICELANDIC LITERATURE

The roots of Norwegian literature reach back more than 1,000 years and become inextricably intertwined with early Icelandic literature. Although a large part of this early literature was composed either in Iceland or elsewhere in Scandinavia by Icelanders, the Norwegian element in it is considerable and indisputable, even though this cannot always be isolated and defined. In many instances, it is obvious that some of the literature derives from a time before the Scandinavian settlement of Iceland in the 9th century. In other cases, it appears that the composers of the works had resided for long periods in the mother country of Norway.

The classical period in Iceland. The best known Icelandic literature belongs to the classical period, roughly equivalent to the early and medieval periods in western European literature. Icelandic manuscripts yield much knowledge of European myth and legend, which is in part common to all Germanic peoples. Stories of the Norse gods and myths—of Odin, god of war; Balder the Beautiful; Thor, god of thunder; and Valhalla, hall of the slain—form the nucleus of early Icelandic literature.

Almost all extant early Scandinavian poetry was recorded in Icelandic manuscripts, although some was clearly composed before the Scandinavian peoples reached Iceland in the late 9th century. Much of the oldest poetry was recorded in the Codex Regius manuscript, which contained the *Sæmundar Edda* (*c.* 1270), commonly designated by scholars as the *Poetic Edda*, or *Elder Edda*. The poetry is sometimes called Eddaic and falls into two sections: heroic lays, which, broadly speaking, dealt with the world of men; and mythological lays, which dealt with the world of the gods. Eddaic poetry

The heroic lays. The heroic lays followed the mytho-

logical in the Codex Regius and were probably the earlier of the two. Many of the legends on which they were based originated in Germany or even among the Goths. Oldest of all was perhaps the *Hamdismál* ("Lay of Hamdir"), which forcefully expressed the heroic ideals of Germanic tribal life. The story closely resembled one told by Jordanes, a Gothic historian of the mid-6th century, and his account suggested that his source was an even earlier poem about Hamdir. Another of the older lays in the *Poetic Edda* was the *Atlakvida* ("Lay of Atli"), which referred to events that took place in 5th-century western Germany, Atli (or Attila) being king of the Huns from 434 to 453. Nearly all heroic lays were associated with the story of Sigurd (or Siegfried), the valiant hero, and his ill-fated love for Brunhild, who, too, figured to varying extent in different lays. Many scholars hold that the lays concerned with the spiritual conflict of the heroines Brunhild and Gudrun, which tend to be romantic and sentimental, were later compositions than the austere heroic lays. The *Poetic Edda* contained only a small portion of the poetry known in Iceland in the Middle Ages and now lost. Fragments of ancient lays appeared in 13th- and 14th-century sagas such as the *Hlödskvida* ("Lay of Hlöd") in the *Heidreks saga*, as did mention of Danish and Swedish heroes in some fragments that must also have been known to the author of the Old English epic poem *Beowulf*.

The mythological lays. Mythological lays about the Norse gods made up the first half of the *Poetic Edda*. It is unlikely that any of these originated outside Norway, Iceland, and Norse colonies in the British Isles. The *Völuspá* ("Sibyl's Prophecy") was a striking poem on the history of the world of gods, men, and monsters, from the beginning until the "twilight of the gods." Many passages in the poem are obscure, but most modern scholars agree that it was composed in Iceland about the year 1000, when the people were turning from the old religion to the new. An interesting story of the gods was told in the *Skírnismál* ("Words of Skírnir"): sitting in "Gate Tower," throne of Odin, the god Freyr, lord of the world, gazes into the world of giants and falls in love with a giant maiden; to win her, he sends his messenger Skírnir, who first offers gifts and then threatens the maiden until she agrees to make a tryst with Freyr. Scholars have seen an ancient fertility myth in this story, and it was certainly one of the older mythological poems in the *Poetic Edda* and probably originated in Norway before Iceland was settled by Norwegians.

Didactic element in Eddaic poetry

The mythological poems so far mentioned were all narrative, but many of those in the *Poetic Edda* were didactic. The *Hávamál* ("Words of the High One"; *i.e.*, Odin) consisted of fragments of at least six poems. In the first section, the god speaks of relations between man and man and lays down rules of social conduct; in other sections he discourses on relations between men and women and tells how love of women may be lost or won; the last two sections are about runes and magic power. Most of the poems were probably composed in Norway in the 9th and 10th centuries. Another didactic poem, the *Vafþrúdnir* ("Words of Vafthrúdnir"), related a contest between Odin and a giant.

Some important mythological lays appeared in other manuscripts. *Baldrs draumar* ("Balder's Dreams") described how the god Balder dreamed that his life was threatened and how his father, Odin, rode to the grave of a prophetess to force her to reveal the fate in store for Balder.

The Eddaic verse forms. Three metres are commonly distinguished in Eddaic poetry: the epic measure, the speech measure, and the song measure. Most narrative poems were in the first measure, which consisted of short lines of two beats joined in pairs by alliteration. The number of weakly stressed syllables might vary, but the total number of syllables in the line was rarely fewer than four. In these respects it resembled the measure used by Anglo-Saxon and early Germanic poets. The speech measure used in the *Atlamál* ("Words of Atli") differed little from the epic measure, though its lines usually had a greater number of weakly stressed syllables. The song measure was the most irregular of the Eddaic verse forms. It was chiefly in didactic poems and generally consisted of strophes of six lines divided into half strophes of three lines.

Skaldic verse. Norwegians and Icelanders of the 9th to the 13th century also composed skaldic poetry (from the Icelandic word *skáld*, "poet"). It was not composed in the free variable metres of the *Poetic Edda* but was strictly syllabic: every syllable had to be counted and every line had to end in a given form. Like Eddaic lines, the skaldic lines were joined in pairs by alliteration, often using internal rhyme or consonance; but this poetry differed in syntax and choice of expression. Word order is freer than in Eddaic poetry, and a highly specialized poetic vocabulary employed periphrases, or kennings, of such complexity that the poetry resembles riddles. Little is known about skaldic verse forms, but they are thought to have been developed in Norway during the 9th century and could have been influenced by the forms and diction of Irish poets of the period. The earliest known poet was Bragi the Old, who probably wrote in Norway in the latter half of the 9th century. Harald I (died *c.* 940) of Norway was eulogized by several poets, among them Thörbjörn Hornklofi, whose poem the *Haraldskvaedi* ("Lay of Harald") was partly Eddaic and partly skaldic in style.

Characteristics of skaldic poetry

The distinction between Icelandic and Norwegian literature at this period is difficult to make. Skaldic verse seems to have originated in Norway and to have been developed by Icelandic poets who either, like Egill Skallagrimsson, spent much time in Norway or wrote in praise of Norwegian kings, as did Sigvatr, counsellor and court poet of Olaf II of Norway. Although its complexity means that skaldic poetry is now less appreciated than it deserves, the orally transmitted poems of the 10th and 11th centuries were valuable sources for Icelandic historians in the following centuries.

Prose. Iceland's adoption of Christianity in 1000 opened the way for powerful influences from western Europe. Missionaries taught Icelanders the Latin alphabet, and they soon began to study in the great schools of Europe. One of the first was Ísleifr, who after being educated and ordained as a priest was consecrated bishop. His school at Skálholt in southern Iceland was for many centuries the chief bishopric and a main centre of learning. The earliest remembered historian was Saemundr the Wise, but Ari Thorgilsson is regarded as the father of historiography in the vernacular. A short history, *Íslendingabók* (or *Libellus Islandorum*, *c.* 1125; *The Book of the Icelanders*), and the more detailed *Landnámabók* ("Book of Settlements") are associated with his name. Extant works of the period are few or anonymous. Annals of contemporary events date from the 13th century and the oldest religious manuscripts, consisting of homilies and saints' lives, from *c.* 1150. Larger collections of religious literature appeared in late 12th- and early 13th-century manuscripts. As elsewhere, the most popular books were often lives of the Apostles and saints.

Influence of Christian missionaries on Iceland

The sagas. The word saga is used in Icelandic for any kind of story or history, whether written or oral. In English it is used to refer to the biographies of a hero or group of heroes written in Iceland between the 12th and 15th centuries. These heroes were most often kings of Norway, early founders of Iceland, or legendary Germanic figures of the 4th to the 8th century. The oldest saga is the fragmentary *Oldest Óláfs saga helga* ("First Saga of St. Olaf"), written about 1180. In form it is a hagiographic narrative, laying emphasis on miracles worked through the agency of the saint. It was probably written in the monastery of Thingeyrar, which played an important part in cultural life in the late 12th and early 13th centuries.

Several sagas about King Olaf Tryggvason, at whose instigation the Icelanders adopted Christianity, were also written at Thingeyrar, where the work of the monks was fanciful rather than realistic. A more critical style of history was established in the south by Saemundr and Ari, and several notable works were written at Skálholt or nearby in the 13th century, such as the *Hungrvaka* ("The Appetizer"), a short history of the bishops of Skálholt from Ísleifr to Kloengr. In the late 12th century several short histories of Norwegian kings were brought from Norway to Iceland, where they influenced Icelandic historians. The *Ágrip*, a summary of the histories, or sagas,

of Norwegian kings, written in the vernacular in Norway, was particularly influential. The *Fagrskinna* ("Fine Skin") covered the same period in more detail, while the *Morkinskinna* ("Rotten Skin"), probably written earlier, covered the period from Magnús the Good (1035–47) to the late 12th century.

The role of Snorri Sturluson

Snorri Sturluson wrote many kinds of works and played an important role in political wrangles in his time. Among works ascribed to him was the *Snorra Edda* (*c.* 1225), a handbook of prosody and poetic diction commonly referred to as the *Prose Edda*, or *Younger Edda*. He twice visited Norway, and a large part of his work consisted of lives of its early kings: he combined his *Ólafs saga* with lives of other Norwegian kings to form the *Heimskringla* (*c.* 1220; "Orb of the World"). The value of these as historical sources has long been debated. Snorri was certainly well read in vernacular history and attempted to write faithful accounts of what he had read in earlier records. He did not aim to write scientific history; his work was creative and therefore portrayed his heroes imaginatively. The stirring *Egils saga* (on the skald Egill Skallagrimsson) is attributed to Snorri.

The Icelanders', or family, sagas. These sagas were about heroes who had supposedly lived in the 10th and 11th centuries. Their origins are unclear, and it is debatable whether they were faithful records of history. One theory is that they were composed in the 11th century and transmitted orally until written down in the 13th century; though researchers now reject this view, it is true that the sagas owed much to oral tales and the tradition of oral verse. Their historicity is difficult to verify, since their content and form were shaped both by the sources used and by the author's intentions.

It is also difficult to determine the date of many of the sagas. The obviously early works were somewhat crudely structured and expressed Norse ideals of loyalty and heroism. The *Gísla saga*, written before the middle of the 13th century, showed a development of artistic skill and contained rich descriptions of nature and verses of considerable beauty and tragic feeling. The *Laxdaela saga* ("Saga of the Men of Laxárdal"), written a few years later, was a delicately worked tragedy in which the author showed an unusual appreciation of visual beauty. One work that was clearly the author's creation was the *Hrafnkels saga Freysgoda* ("Saga of Hrafnkell, Freyr's Priest"): despite realistic detail, the saga contained little historical fact. As the century progressed, a taste for fantastic and romantic elements grew. The *Grettis saga* ("Saga of Grettir the Strong") included several motifs from folklore and portrayed a hero fighting against trolls and ghosts.

The saga of Njáll and Gunnar

The greatest of Icelanders' sagas, the *Njáls saga*, had in fact two heroes, Njáll and Gunnar. Gunnar is young and inexperienced and Njáll is a wise and prudent man endowed with prophetic gifts; he embodies traditional Norse ideals of loyalty and bravery, yet faces his death by burning with the resignation of a Christian martyr.

The heroic sagas. The fantastic element was further developed in the *fornaldar sögur*, literally "the sagas of antiquity," whose heroes were supposed to have lived in Scandinavia and Germany before Iceland was settled. The best known, the *Völsunga saga* (*c.* 1270), retold in prose stories from heroic lays of Sigurd, the Burgundians, and Jörmunrekr, and the *Hrólfs saga kraka* (*c.* 1280–1350) incorporated ancient traditions about Danish and Swedish heroes who also appeared in the Old English poems "Widsith" and *Beowulf*.

Many of the works on contemporary history were combined about 1300 in the *Sturlunga saga*, including the *Íslendinga saga* by Sturla Thórdarson.

Translations from Latin. A quantity of secular literature was translated from Latin between the 12th and 14th centuries. The "Prophecies of Merlin," already translated in verse by a Thingeyrar monk, were combined with a complete translation of Geoffrey of Monmouth's history and titled *Breta sögur* ("Stories of the Britons"). In one 14th-century manuscript this was preceded by the *Trójumanna saga* ("Story of the Trojans"), translated from Dares Phrygius. A Norwegian translation of the Bible was begun in the reign of Haakon V Magnusson (1299–1319).

Romances. Romances were also translated or adapted from continental romances. Interest in romance began in Norway and soon took root in Iceland. The earliest romance was probably the *Tristrams saga* (1226), derived from the Anglo-Norman poet Thomas. This was followed by the *Karlamagnús saga* ("Saga of Charlemagne"), a collection of prose renderings of French chansons de geste, including a Norse version of the *Chanson de Roland*. Romances in Icelandic were numerous, and their effect on the style of later writers is evident in such sagas as the *Laxdaela saga* and *Grettis saga*.

Post-classical literature in Iceland. In the period following the classical age, little was written that attracted attention outside Iceland. Realism and detached objectivity declined, and sentimentality and fantasy gained the upper hand. The decline in literary standards is sometimes attributed to Iceland's loss of independence in 1262 and the changes that followed. Interest in earlier manuscripts continued, and many 14th- and 15th-century manuscript collections of 13th-century material were made. The most beautiful of all Icelandic manuscripts, the *Flateyjarbók* (*c.* 1390), included versions of sagas of Olaf Tryggvason and St. Olaf, together with texts from other sagas or about heroes associated with Iceland.

Decline in literary standards

Prose. Prose literature of the 14th century included several sagas. Among them were the *Finnboga saga ramma* ("Saga of Finnbogi the Strong"), about a 10th-century hero, and another telling the love story of its hero Víglundr. Sagas about bishops, already a theme in the 13th century, became more numerous, as did lives of foreign saints. A large collection of exempla (moral tales) was also made, each short tale illustrating some moral precept.

Poetry. Much poetry was written up to the time of the Reformation, and many new forms were devised. The best poems were religious pieces, in honour of the Virgin, the Apostles, or other saints. The well-known *Lilja* (*c.* 1350; "The Lily") by Eysteinn Ásgrímsson, a monk from Thykkvabaer, gave an account of the fall of Satan, the creation, the first sin, and the birth, life, and Passion of Christ. The term *rímur*—rhymes—is used of the narrative poetry developed after 1500 that consisted of mainly four-line strophes: the lines had end rhyme. The metrical forms, although apparently derived from Latin hymns, inherited the alliterative system of earlier poetry. Ballads written in Icelandic never attained the popularity of Danish ballads in Denmark nor achieved the high standard of the Norwegian *Draumkvaede* ("Dream Ballad"). Most of those preserved dated from the 14th to the 16th century and were free translations of Danish and Norwegian originals.

Ásgrímsson's *Lilja*

SWEDISH LITERATURE

Swedish literature proper began in the late Middle Ages when, after a long period of linguistic change, Old Swedish emerged as a separate language. The foundations of a native literature were established in the 13th century. The oldest extant manuscript in Old Swedish was the *Västgötalagan* ("Law of West Gotland"), part of a legal code compiled in the 1220s. These legal documents often employed concrete images, alliteration, and a solemn prose rhythm suited to their proclamatory nature.

The poetry of chivalry was first represented in *Eufemiavisorna* ("The Songs of Euphemia"), written in doggerel between 1303 and 1312, which included a translation of Chrétien de Troyes' romance *Yvain*. Anonymous ballads probably dating from the 14th and 15th centuries also reflected a new interest in romance. These ballads, though mostly derived from foreign sources and combining the imported ideals of courtly love with native, pagan themes and historical events, formed the most accessible genre of what can be called Swedish medieval literature.

Early Swedish ballads

DANISH LITERATURE

Denmark's first literature appeared in the runic inscriptions scratched on stone or carved in metal, mainly epitaphs of warriors, kings, and priests that occasionally had short, unrhymed alliterative verses in the Viking spirit. Runic inscriptions were used in Denmark from about 250, but most of those preserved date from 800 to 1100. With the introduction of Christianity, Latin became the

predominant literary language, and Denmark's first important contribution to world literature, Saxo Grammaticus' *Gesta Danorum* (written between 1185 and 1222; "The Deeds of the Danes"), which contained, for example, the Hamlet story, was written in Latin. The medieval ballads of Denmark are among the most important in Europe; 539 are known in more than 3,000 versions, but nearly all were written down after the end of the Middle Ages, the first printed edition appearing in 1591.

The 16th century

THE IMPACT OF THE REFORMATION ON SWEDISH LETTERS

Two dates mark the beginning of modern Swedish history: 1523—the breach with Denmark and Gustav I Vasa's accession; and 1527—the breach with Rome and the establishment of a national Lutheran Church. The political revolution that eventually brought Sweden to the position of a European power had no considerable effect on literature until a century later, but the Reformation wholly dominated Swedish letters in the 1500s.

The Swedish translation of the Bible

The most important literary event of this period was the translation of the Bible in 1541, which inaugurated modern Swedish and provided an inexhaustible source for poets of subsequent times. Closely involved in the Bible translation were the apostles of the Swedish Reformation, Olaus Petri and his brother Laurentius. Olaus Petri's vigorous approach was revealed in his published sermons and in a Swedish chronicle, the first historical Swedish work based on critical research. Olaus Petri may also have written the biblical *Tobie comedia* (published 1550), the first complete extant Swedish play.

As a consequence of the Reformation, two of Sweden's most distinguished scholars of the period, Johannes Magnus and his brother Olaus, were driven into exile. In his history of all the kings of the Goths and Swedes, Johannes provided Sweden with a number of valiant kings unknown to critical historians. Olaus wrote the first geographical and ethnographical account of Scandinavia, *Historia de gentibus septentrionalibus* (1555; "History of the Northern Peoples").

DEVELOPMENTS IN DANISH LITERATURE

In 1536 the Lutheran Reformation was carried through in Denmark, and the beginning of the 16th century was characterized by many pamphlets for or against the Roman Catholic Church. European humanism and the Renaissance made their influence felt also in Denmark, where Christiern Pedersen was the most prominent humanist who supported the Reformation. He edited *Gesta Danorum* by the 13th-century historian Saxo Grammaticus, translated the New Testament, adapted Martin Luther's pamphlets into Danish, and participated in a translation of the Bible (1550). Poul Helgesen was the most gifted opponent of the Lutheran Reformation and Hans Tausen its most talented spokesman. The *Visitation Book* by the Lutheran bishop Peder Palladius is an important literary document. The two most important historians were Anders Sørensen Vedel and Arild Huitfeldt.

Sixteenth-century Danish poetry was religious or polemical, with fine love poetry and hymns. The earliest plays date from the beginning of the century. The most important playwright of the period was Hieronymus Justesen Ranch, whose farce *Karrig Nidding* ("The Miserly Rascal") was his best play.

ICELANDIC LEARNING AND LITERATURE

The chief political figure and poet of the Reformation was Jón Arason, last Catholic bishop of Hólar, beheaded in 1550. By his life Jon showed that he was a Viking as well as a martyr, although most of his surviving poetry is religious.

The impact of the Reformation

The effect of the Reformation on Icelandic learning and literature was that Catholic poetry was discarded and attempts were made by the first Lutheran bishops to replace it with hymns poorly translated from Danish and German. Lutheran teachers instructed the people in Protestant dogma, and several translations of sermons and books of instruction by German Lutherans were printed in Icelandic from as early as 1540. Gudbrandur Thorláksson was the most energetic of the Lutheran teachers. In translating the Bible he used earlier Icelandic versions of some books of the Old Testament and Oddur Gottskálksson's Icelandic translation of the New Testament. In his psalmbook he showed appreciation of Icelandic poetic tradition and adhered to Icelandic alliteration and form.

The 17th century

SWEDISH POETRY AND PROSE

In the first half of the 17th century, Swedish literature remained limited in scope and quantity. A unique contribution, however, was made by Lars Wivallius, whose lyrics revealed a feeling for nature new to Swedish poetry. With its intervention in the Thirty Years' War, Sweden established itself as a European power, and this led to a development of national pride and culture, as revealed in literature of this epoch. The outstanding work was the allegorical epic *Hercules* (1658) by Georg Stiernhielm, which reflected many of the social and political problems of the time. Stiernhielm's followers included the two brothers Columbus, one of whom, Samuel, wrote *Odae sueticae* (1674; "Swedish Odes") and the prose *Mål-roo eller roomål*, a charming collection of anecdotes that illumine Stiernhielm's character. A rival to Stiernhielm was the unidentified "Skogekär Bärgbo," whose *Wenerid* (1680) was the first sonnet cycle in Swedish.

Georg Stiernhielm and his followers

Stiernhielm aimed at an integration of Sweden's cultural heritage with the accepted ideals of continental classicism. His *Hercules* is full of old Swedish words that he was eager to revive. Columbus also demanded a more vigorous, flexible language as did "Skogekär Bärgbo" in *Thet swenska språkets klagemål* (1658; "The Lament of the Swedish Language"). National pride and religious feeling are combined in the works of the bishops Haquin Spegel and Jesper Swedberg, father of the Swedish mystic Emanuel Swedenborg. Spegel contributed to Swedberg's new hymnbook of 1695, which became the poetry book of the Swedish people and was of lasting influence. Even Lucidor was represented in it, giving intense expression to the contrasting moods of the period: in his love songs and, above all, in his drinking songs, he was as pagan and reckless as he was devout in his hymns and funeral poems.

At Uppsala, meanwhile, the scholar Petrus Lagerlöf attempted to impose purer classical standards on native literature, and Olof Verelius edited and translated Icelandic sagas. It was Olof Rudbeck, however, who became interested in Verelius' work and developed a theory that Sweden was the lost Atlantis and had been the cradle of Western civilization. He proposed this idea in *Atland eller Manheim* (1679–1702), which, translated into Latin as *Atlantica*, attained European fame.

Baroque and classicist tendencies ran parallel in late 17th-century Swedish literature. Gunno Eurelius (Gunno Dahlstierna) wrote an elaborate epic, *Kungaskald* ("Hymn to the King"), for King Charles XI's funeral in 1697. Simpler in style was Johan Runius, who expressed a Christian stoicism of the kind found among Swedes during the disastrous early decades of the 18th century. Jacob Frese was a gentler and more intimate poet; his lyrics and hymns contained some of the emotional pietism that became a feature of 18th-century thought.

THE LITERARY RENAISSANCE IN DENMARK

The literary Renaissance reached Denmark in the 1600s, giving rise to a strict adherence to classical patterns and blind belief in authority in political, religious, and literary matters. In religious literature Latin dogmatics and pamphlets reflecting the superstitions of the century were dominant. It was, however, a great era of scholarship. Ole Worm is famous for his book on the runic inscriptions, *Monumenta Danica* (1643). Thormod Torfaeus and Árni Magnússon introduced the study of Old Norse literature; Peder Hansen Resen edited and translated some of the poetry of the Old Norse *Edda;* and Erik Pontoppidan and Peder Syv introduced the linguistic study of Danish.

The revival of interest in Scandinavian antiquity

Danish poetry in the 17th century tended to follow the

classics slavishly, and the favourite forms were the hexameter, the Alexandrine, and the sonnet. Simplicity is deliberately avoided; the style is precious; allegories, euphemisms, and metaphors abound. Anders Arrebo translated the Psalms and wrote *Hexaëmeron* (1661), a Danish version of the 16th-century French poet Guillaume du Bartas' *La Semaine*. The century was rich in occasional poetry; didactic and pastoral poems were also common. Anders Bording, an interesting exponent of Danish Baroque poetry, was also the founder of the first Danish newspaper, *Den danske Mercurius* (from 1666), in which the news appeared in rhymed Alexandrines. The only truly great poet was Thomas Kingo, a supreme master in almost every kind of poetry. His hymns reflect a violent, passionate character, worldly and yet deeply religious.

Characteristic poetic forms

Of special interest among Danish works of the 17th century were the memoirs of Leonora Christina, daughter of Christian IV, a fascinating document about her 20 years' imprisonment in the Blue Tower of Copenhagen.

RENEWED LITERARY ACTIVITY IN NORWAY

Political union between Denmark and Norway started in 1380, and the Danish language eventually became the official and the literary medium. Copenhagen, with its university, established itself as the cultural capital of the two countries. Not until after the Reformation were there signs of renewed literary activity in Norway itself; *e.g.*, in the nostalgic apologia for Norway, *Om Norgis rige* ("Concerning the Kingdom of Norway"), written in 1567 by Absalon Pederssøn Beyer. The most original and most conspicuously Norwegian writer of this age was Petter Dass, whose *Nordlands trompet* (*The Trumpet of Nordland*) gives a lively picture in verse of the life of a clergyman; although probably completed before the turn of the century, this work was not printed until 1739.

ICELANDIC LETTERS

In Iceland the foremost poet of the 17th century was Hallgrímur Pétursson, a Lutheran pastor who struggled against poverty and ill health. His *Passíusálmar* (1666; "Hymns on the Passion") is among the most popular books in Iceland. Another interesting poet was Stefán Ólafsson, remembered for both religious and secular works, the latter notable for exuberantly humorous portrayals of contemporaries and satirical observations of manners and customs.

As in other countries, interest in antiquity was stirred during the 17th century, and modern learning may be said to date from that period. Arngrímur Jónsson called the attention of Danish and Swedish scholars to Icelandic traditions and literature in a series of works in Latin, some containing abstracts of sagas now lost. Later in the century Árni Magnússon systematically collected the early Icelandic manuscripts.

The 18th century

SWEDISH CLASSICISM AND ENLIGHTENMENT

After the death of Charles XII (1718) and the collapse of his empire, a utilitarian attitude to life and letters gradually developed in Sweden. Olof von Dalin was the outstanding popularizer of the new ideas of the French and English Enlightenment. Educated at Lund, he later went to Stockholm and began to publish, anonymously, *Then swänska Argus* (1732–34; "The Swedish Argus"), a weekly periodical modelled on that of the Englishman Joseph Addison. One of the first serious journalistic ventures in Sweden, it marked the beginning of a new era, in which orthodoxy gave way to Skepticism and Enlightenment, Baroque to Classicism, and German influence to English and French; at this time the middle class began gradually to take over the function of chief upholder of literature. In *Argus* Dalin ridiculed the foibles of the capital and in *Sagan om hästen* (1740; "The Story of the Horse") he showed himself a master of allegorical satire. He also produced some pseudo-Classicist plays that, like many dramatic ventures of the early and mid-18th century, are academic and lifeless. The one notable exception is *Den Svenska prätthöken* (1740; "The Swedish Fop"), a comedy by Count Carl Gyllenborg.

English and French influence

With the second phase of the Enlightenment, marked by the influence of Rousseau, are associated Hedvig Charlotta Nordenflycht, the epicurean Gustav Philip Creutz, and his stoic friend Gustaf Fredrik Gyllenborg. In *Den Sörgande turturdufwan* (1743; "The Sorrowing Turtledove"), Fru Nordenflycht laments the death of her husband in highly personal lyrics. Creutz was a more sophisticated personality. He wrote little, but his few writings, of which the pastoral *Atis och Camilla* (1762) is the most important, reveal a mastery of form and versification.

Prose—particularly the novel—developed more slowly. The first genuine novel, *Adalrik och Giöthildas äfventyr* (1742–44; "The Adventures of Adalrik and Giöthilden"), by Jacob Mörk and Anders Törngren, shows the influence of the Icelandic sagas. Only two 18th-century Swedish writers were of European reputation, and both were scientists: Carl von Linné (Linnaeus) and Emanuel Swedenborg.

The first Swedish novel

The Gustavian period takes its name from King Gustav III (1746–92), a brilliant man and a patron of art and letters. He was especially interested in drama and opera and, thanks to his patronage, a proper theatrical tradition developed. Gustav himself sketched out some works, the best of which was a historic opera, *Gustaf Vasa,* which was finished in collaboration between Johan Henrik Kellgren and the composer J.G. Naumann. Kellgren, an academic poet and arbiter of taste, ruled that Swedish literature should be modelled on Classicist French patterns, but, beginning as a Rationalist and satirist after the fashion of Voltaire, he reluctantly accepted pre-Romantic ideas later. In *Stockholmsposten,* the main organ of literary opinion in the capital, Kellgren used his polemical wit against Thomas Thorild, a truculent champion of individual genius. After Kellgren's death the controversy was carried on by Carl Gustaf af Leopold, who imposed pseudo-Classical standards on the academy and applied them in his own rhetorical odes and tragedies. Johan Gabriel Oxenstierna did his most original work while a diplomat in Vienna; his *Skördarne* (1796; "Harvests") reveals pre-Romantic feeling for the beauty of nature. Bengt Lidner was the chief exponent of pre-Romanticism in poetry. His most successful work was the ode *Grefvinnan Spastaras död* (1783; "The Death of Countess Spastara").

Carl Michael Bellman stands apart from the conflicting ideals of the time. A poet and musician, he combined stylized realism with humour and the most uniquely delicate sense of language and rhythm. He was the greatest Swedish lyricist of the 18th century.

The dissertation *Om upplysning* (1793; "On Enlightenment") by Nils von Rosenstein, the first secretary of the Swedish Academy, expressed the ideals of the Gustavian epoch. Memoirs by G.J. Adlerbeth, G.J. Ehrensvärd, and others evoke the witty but artificial atmosphere of Gustav III's court. Gustav IV, who followed, did not encourage literature; nevertheless, Anna Maria Lenngren wrote some of her best verse satires between 1795 and 1800, many aimed at aristocratic foibles. The sentimental idylls of Frans Mikael Franzén are full of pre-Romantic idealism from German and English sources.

LITERARY ACTIVITY IN DENMARK, NORWAY, AND ICELAND

Denmark. The 18th century was a fertile period in Danish literature. The great name in the first half of the century was that of Ludvig Holberg, a Norwegian by birth. His most important contributions, written for the Danish theatre, which opened in 1722, were 32 comedies of character and manner, including some moral allegories in his old age. His aim was to create a modern Danish literature on European lines and to make people laugh at their own follies. Influenced by English and French thinking, he was a Rationalist and a moderate. He also wrote satire, a mock-heroic poem, and *Nicolai Klimii iter Subterraneum* (Latin, 1741; *Journey of Niels Klim to the World Underground*). His *Moralske tanker* (1744; "Moral Thoughts") and his *Epistler* (1748–54; "Letters") are the finest examples of a Danish political essay form.

The comedies of Ludvig Holberg

Among Holberg's contemporaries the finest lyrical poets are H.A. Brorson, a mystic whose pietist hymns often have a background of personal sorrow or agony; and

Ambrosius Stub, whose poems are mainly religious and moralizing verses, witty epigrams, or drinking songs. A satirist, Christian Falster, was a conservative counterpart to Holberg; Friedrich Eilschov and Jens Schelderup Sneedorff, the latter of whom edited *Den patriotiske Tilskuer* ("The Patriotic Spectator"), a Danish *Spectator*, were both Rationalist disciples of Holberg.

Revival of Danish literature

A significant revival of Danish literature took place toward the end of the century. In 1772 the Norwegian Johan Herman Wessel, one of the greatest humorists to use the Danish language, wrote *Kaerlighed uden strømper* ("Love Without Stockings"), a parody of the Danish imitations of Italian operas and French tragedies that had superseded Holberg's comedies.

The works of Johannes Ewald

At the same time a revival of emotional poetry was taking place, influenced by German and English literature. Johannes Ewald, perhaps Denmark's greatest lyrical poet, was the first to discover the poetic wealth of Scandinavian antiquity in the *Gesta Danorum* of Saxo Grammaticus and in the myths, sagas, and ballads. He wrote verse dramas and deeply personal and descriptive poems. *Fiskerne* (1779; "The Fishermen") was the first serious Danish drama in which ordinary people were treated heroically. His memoirs, *Levnet og meninger* (posthumously published in 1804; "Life and Opinions"), were influenced by Laurence Sterne and Jean-Jacques Rousseau. Jens Baggesen at first imitated the satires of Holberg and Wessel but gradually developed as a poet of distinction. In *Labyrinten* (1792–93; "The Labyrinth"), he described his travels in Europe in the manner of Sterne.

Norway. Several of Denmark's leading writers of the 18th century were of Norwegian birth, preeminently Ludvig Holberg and the members of Det Norske Selskab (the Norwegian Society). Established in Copenhagen in 1772 by a group of resident Norwegians, it looked to French rather than to German and English literature for models. Within Norway itself there was little overt literary activity, though the establishment in 1760 of a Royal Norwegian Society of Learning in Trondheim was evidence that Norway was beginning to assert its cultural aspirations.

Iceland. *Húss-Postilla* (1718–20; "Sermons for the Home"), by Jón Vídalín, bishop of Skálholt, is the best example of early 18th-century prose. Among important later writers, Eggert Ólafsson carried out a comprehensive geographical field survey (published in Danish, 1772; partial Eng. trans.) of Iceland's country and its people. In his poetry he expressed 18th-century Rationalism combined with Romantic patriotism. Jón Thorlákssón, poet and scholar, translated John Milton's *Paradise Lost* and Alexander Pope's *Essay on Man*.

Finnur Jónsson, bishop of Skálholt, wrote *Historia Ecclesiastica Islandiae* (1772–78), which covers the history of Christianity in Iceland. Jón Espólín published *Íslands árbaekur* (1822–55; "Annals of Iceland"), a history of Iceland from 1262.

The 19th century

SWEDISH LITERATURE

Romanticism. Political changes in Sweden up to 1804 meant that ardent nationalism emerged as a characteristic of Swedish Romanticism. The idealism at the core of this movement was laid by the Kantian teaching of Benjamin Höijer and the impact of Friedrich Schiller, Johann Wolfgang von Goethe, and the German Romantics on Swedish literature. Student societies and their periodicals, such as *Polyfem* (1809–12) and *Phosphorus* (1810–13), led the attack on the traditional school. Most gifted of the Forforister, or Phosphorists, Per Daniel Atterbom, wrote a verse "Prolog" (1810) to *Phosphorus* revealing both talent and commitment to Romanticism.

The return to the literary past

Meanwhile, another society, Götiska Förbundet (Gothic Society), advocated, from its start in 1811, that study of the "Gothic" past could morally improve society. One of its members, Esaias Tegnér, wrote a most popular poem, *Frithiofs saga* (1825), based on an Old Norse theme. Tegnér valued old Northern mythology for the patterns he discerned in it—patterns also found in Greek mythology and Romantic metaphysics, in which religion, philosophy, and poetry appeared to be one and the same. Nevertheless, Tegnér's ideals of clarity of thought and formal perfection led him sometimes to side with traditionalists in their struggle against obscurities and formal innovations.

Several leading Romantics were learned men whose poetry strove to embody a philosophical system or an interpretation of history. The most ambitious attempt of this kind was P.D.A. Atterbom's *Lycksalighetens ö* (1824–27; "The Isle of Bliss"), an allegory dealing with adventures of a legendary king, Astolf, and a history of poetry as an illustration of man's alienation from the divine. The greatest poet was perhaps Erik Johan Stagnelius, who held aloof from schools and coteries. The recurrent theme in his *Liljor i Saron* (1821; "Lilies of Sharon") was the lament of the human soul, imprisoned in a world of darkness and sin.

In prose the most complex personality among the later Romantics was a novelist, Carl Jonas Love Almqvist, who combined an extravagant imagination with realism. A master of prose style, he was at his best in the long short story, in which he foreshadowed Strindberg's method of raising problems for debate. The novel was established by Fredrika Bremer, author of *Grannarna* (1837; "The Neighbours"), whose "sketches from ordinary life" appeared from 1828. Sophie von Knorring wrote chiefly about aristocratic families, and Emilie Flygare-Carlén produced stories dealing with west-coast life, including *Rosen på Tistelön* (1842; *The Rose of Tistelön*).

Emergence of Realism and Poetic Realism. Realism made only slow headway in spite of the example of the Finno-Swedish poet Johan Ludvig Runeberg (see below). Literature of the 1840s and 1850s was mainly an aftermath of Romanticism. A movement known as Scandinavism produced a good deal of verse: Carl Vilhelm August Strandberg (pseudonym "Talis Qualis"), fieriest poet of this type, later made excellent translations from Byron. Popular reading was provided by August Blanche in *Bilder ur verkligheten* (1863–65; "Pictures of Real Life"), short stories depicting Stockholm life with humour and vivacity, while Frans Hedberg wrote pompous historical plays.

Scandinavism

Poetic Realism became an official program of the "pseudonym poets" of the 1860s, including Carl David of Wirsén, Edvard Bäckström, Pontus Wikner, and Carl Snoilsky. Only Snoilsky had the temperament and poetic gift needed to carry out the program. Wirsén, on the other hand, as secretary of the Swedish Academy, launched formidable opposition against innovators; and Viktor Rydberg fell between Idealism and Naturalism. His important early work consisted of an ideological novel, *Den siste athenaren* (1859; *The Last Athenian*), and a treatise, *Bibelns låra om Kristus* (1862; "The Teaching of the Bible About Christ"), which prepared the way for scientific Rationalism.

Sources of modern Swedish literature. Four influences combined to free Swedish literature from petrifying conventions: the English writings of Charles Darwin, Herbert Spencer, and John Stuart Mill; the French Naturalism of Émile Zola; the drama of the Norwegians Henrik Ibsen and Bjørnstjerne Bjørnson; and the criticism of the Dane Georg Brandes. The modern literature growing out of this was first and best represented in the work of August Strindberg, Sweden's greatest writer. Modern drama has dated from his play *Mäster Olof* (1872), and the modern novel from *Röda rummet* (1879; *The Red Room*). Strindberg overshadowed all the writers of the 1880s, including Gustaf af Geijerstam, author of *Erik Grane* (1885), Anne Charlotte Edgren-Leffler, and the gifted Victoria Benedictsson; the latter two wrote about the adverse position of women in society. Benedictsson's stories, such as *Från Skåne* (1884; "From Skåne"), revealed the regional character of the new prose literature. Regional poetry was written by Albert Bååth and Ola Hansson, both of Skåne.

August Strindberg

In 1888 Verner von Heidenstam began the reaction against Utilitarianism and Naturalism with a volume of verse, *Vallfart och vandringsår* ("Pilgrimage and Wander Years"). His later poetry and historical tales won him the Nobel Prize for Literature in 1916. Oscar Levertin, stimulated by Heidenstam's example, wrote poetry full of colour and lore of the past and as a critic was influential in molding contemporary taste. Gustaf Fröding was also

influenced by Heidenstam, and his verse constantly mingles the melancholy and gay. Regionalism entered Neoromantic poetry with Fröding, who was from Värmland, and with the work of Erik Axel Karlfeldt the province of Dalarna came into its own. Karlfeldt's mature poetry won him the Nobel Prize in 1931.

Selma Lagerlöf

Meanwhile, Selma Lagerlöf, the first Swede to win a Nobel Prize for Literature (1909), had developed the prose tale; her long series of novels and short stories, beginning with *Gösta Berlings saga* (1891), reached an international public through translation. Per Hallström was a more skillful writer of short stories than of novels. Romantic, too, in his love for the skerries (rocky isles) was Albert Engström, a great humorist.

Finno-Swedish literature. A significant literature in the Swedish language developed in Finland during the 1800s. Its emergence can be traced to the works of Johan Ludvig Runeberg. His epic poems, *Elgskyttarne* (1832; "The Moose Hunters") and *Hanna* (1836), won him a place in Swedish letters. Notable, too, were the writings of Zacharias Topelius, which contributed to the development of the Finnish historical novel. Topelius is perhaps best remembered for *Fältskärns berättelser* (1853–67; *The King's Ring and the Surgeon's Stories*), a romanticized account of 17th- and 18th-century Finno-Swedish history.

NORWEGIAN LITERATURE

The Age of Wergeland. After 1814 a new, exciting, and difficult age began for Norway: an opportunity seemed to be offered to develop an independent Norwegian culture and way of life, but there were deep differences of opinion as to how this could best be achieved. A poet and critic, Johan Sebastien Welhaven was chief representative of those who insisted that the existing Danish element in the culture should not be neglected. Henrik Wergeland was a spokesman for those whose nationalistic pride led them, on the other hand, to demand a complete break with Denmark. Welhaven stood for a coolly intellectual approach, for restraint and control, and for conscious artistry, as his own sonnet cycle *Norges daemring* (1834; "The Dawn of Norway") exemplifies. Wergeland was more passionate and revolutionary, and his enormous epic, *Skabelsen, mennesket og messias* (1830; "Creation, Humanity and Messiah"), typified the spirit he admired.

Wergeland dominated the age as poet, orator, and social reformer, and the clash between him and Welhaven and between the two factions associated with them—the "patriots" and the "intelligentsia"—began an ideological conflict that has continued to persist in modified forms.

National Romanticism. The literature of the mid-19th century, known as Norway's "national Romanticism," continued to reflect the country's larger aspirations. The compilation and publication, between 1841 and 1844, of *Norske folkeeventyr* ("Norwegian Folk Tales") by Peter Christen Asbjørnsen and Jørgen Engebretsen Moe—and the 1853 collection by Magnus Brostrup Landstad, *Norske folkeviser* ("Norwegian Folk Ballads")—indicated a lively interest in the past, as did Peter Andreas Munch's eight-volume history of the Norwegian people (1857–63). Ivar Aasen was the creative spirit behind the Landsmål movement to establish a literary language based on rural dialects linked with Old Norse. Many publications of these years, including earlier works of Ibsen and Bjørnson, turned consciously to Norway's heroic past and its peasants. To these years belonged also the lyric poetry of Aasmund Olafsson Vinje, founder of the periodical *Dølen,* who adopted Nynorsk (New Norwegian) as his literary language.

In 1855 Camilla Collett, Wergeland's sister, published *Amtmandens døttre* ("The Governor's Daughters"), which, by considering the place of women in society, marked a beginning of a trend that, encouraged by the immensely influential Danish critic Georg Brandes, culminated in the 1870s and the '80s in the realistic "problem" literature of Ibsen, Bjørnson, and their contemporaries. *Samfundets støtter* (1877; *Pillars of Society*) was the first of a succession of problem dramas by Ibsen to win him worldwide fame. By then he had already written two verse dramas, *Brand* (1866) and *Peer Gynt* (1867), and his long "double drama" *Kejser og Galilaeer* (1873; *The Emperor and the Galilean,* 1876). The first substantial drama of this type by Bjørnson was *En fallit* (1875; *The Bankrupt*). Although never the world figure that Ibsen became, Bjørnson was a leading personality of his age in Norway, as novelist, dramatist, and lyric poet and in public affairs.

The problem dramas of Ibsen and Bjørnson

The novelists Jonas Lie and Alexander Kielland, together with Ibsen and Bjørnson, were the major figures of modern Norwegian literature and were responsible for a remarkably large body of important work between 1870 and 1884, as the following titles illustrate: Ibsen's works *Et dukkehjem* (*A Doll's House*), *Gengangere* (*Ghosts*), *En folkefiende* (*An Enemy of the People*), and *Vildanden* (*The Wild Duck*); Bjørnson's dramas *Det ny system* (*The New System*), *En handske* (*A Gauntlet*), and *Over aevne* (*Beyond Human Power I*) and his novel *Det flager i byen og på havnen* (*The Heritage of the Kurts*); Lie's novels *Gaa Paa!* ("Go Ahead!"), *Livsslaven* (1883; "The Life Convict"; Eng. trans. *One of Life's Slaves,* 1895), and *Familjen paa Gilje* (*The Family at Gilje*); and Kielland's *Skipper Worse* (1882), *Gift* (1883; "Poison"), and *Fortuna* (1884; *Professor Lovdahl*). The foremost stylist of his age, Kielland was an elegant, witty novelist with a strong social conscience and an active reforming zeal stemming from an admiration for John Stuart Mill.

Major figures of modern Norwegian literature

The literature of the 1870s emphasized individual development and expression in keeping with the optimistic attitude of the times to social change and improvement. In the following decade, growing skepticism and disillusionment made writers more bitter in their attacks on "established" social institutions. The publication of *Fra Kristiania-Bohêmen* ("From the Christiania Bohemia") in 1885 by Hans Henrik Jaeger created, by its seeming advocacy of sexual license, a public scandal. The most extreme exponent of Naturalism was Amalie Skram, especially in a four-volume novel, *Hellemyrsfolket* (1887–98; "The People of Hellemyr"). Arne Garborg, poet, novelist, dramatist, and critic, was a much superior writer whose work reflected successive movements of Romanticism, Realism, Naturalism, and Neoromanticism. His wider reputation was first established with a novel, *Bondestudentar* (1883; "Peasant Students"), but perhaps his greatest achievement was the poem cycle *Haugtussa* (1895).

DANISH LITERATURE

The Romantic period. The Romantic movement came to Denmark from Germany, inspired partly by the German Jena Romantics and partly by the Neoclassicism of Goethe and Schiller. Friedrich Schelling's philosophy was interpreted in Denmark by the Norwegian Henrik Steffens, but the leading Danish Romantics gave it a form very different from the original. The leader of the Romantic movement in Denmark was Adam Oehlenschläger, whose unparalleled versatility in poetry, drama, and prose showed the influence of certain works of Goethe and Schiller and of German Romanticism. His plays *Sanct hansaften-spil* (1802; "Play for Midsummer Eve") and *Aladdin; Hakon Jarl* (1857), one of his many Northern tragedies; and a cycle of dramatic poems, *Helge* (1814), were outstanding. The popular and historical songs and hymns of the poet N.F.S. Grundtvig, as well as his personal poetry, have given him a lasting place in Danish literature. Sharing the Romantic enthusiasm for antiquities of Scandinavia, he translated the 13th-century historians Saxo Grammaticus and Snorri Sturluson (see above) and translated *Beowulf* even before it had appeared in English. Bernhard Severin Ingemann wrote historical novels and a poetic cycle, *Holger Danske* (1837; "Holger the Dane"), around the themes of chivalry and nationalism as well as his unsophisticated *Morgen og aftensange* (1837–39; "Morning and Evening Songs"). Johannes Carsten Hauch wrote tragic and philosophic dramas, novels, and contemplative poetry.

The works of Adam Oehlenschläger

Romantic Realism. New elements of reason and realism appeared after the first quarter of the century in the works of Poul Møller, who wrote the first Danish novel on contemporary life, *En dansk students eventyr* (1824; "The Adventures of a Danish Student"), and dramatic poems and fables, sometimes showing personal disillusionment, and of Steen Steensen Blicher, who in *Traekfuglene* (1838; "The Birds of Passage") interpreted human nature

with sad resignation. Some of his best poems were in the Jutland dialect. His many *noveller,* or short stories, beginning in 1824 with the masterly "En landsbydegns dagbog" ("The Journal of a Parish Clerk"), struck notes varying from sorrow and resignation to humour and irony.

Minor writers of the same period were Thomasine Gyllembourg-Ehrensvärd, whose novel *En hverdagshistorie* (1828; "A Story of Everyday Life") was much admired; Andreas de Saint-Aubin, who wrote novels under the nom de plume of "Carl Bernhard"; and Carl Bagger, whose novel *Min broders levned* (1835; "My Brother's Life") shocked the literary world by its bold realism.

Basic characteristics

Poetic Realism. About 1830, early Romanticism gave way to a less naive poetic realism, more contemplative and more concerned with form than with content. Johan Ludvig Heiberg, who led this movement, attempted to revivify Danish drama by importing French vaudeville, and in his serious romantic plays *Elverhøj* (1828; "The Elfinhill") and *Syvsoverdag* (1840; "Day of the Seven Sleepers") he juxtaposed poetic and pedestrian reality. His finest achievement was a verse comedy, *En sjael efter døden* (1841; "A Soul After Death"). He was the leading literary critic of his time, profoundly influenced by the philosophy of G.W.F. Hegel. Henrik Hertz also regarded the perfection of poetic form as more important than its content, as was clearly expressed in *Gjenganger-breve* (1830; "Letters of a Ghost"). He also wrote comedies and serious Romantic plays, including *Kong Renés datter* (1845; *King René's Daughter*).

An upsurge of interest in lyrical poetry occurred in the 1830s and 1840s, led by poets concerned with the aesthetic treatment of love and nature. Christian Winther, best known for a long verse novel, *Hjortens flugt* (1885; "The Flight of the Stag"), sang the praises of his native island, Zealand, and of woman. Ludvig Bødtcher wrote delicate and sensitive poetry, some of which was inspired by the Italian scene. Emil Aarestrup treated erotic themes. Frederik Paludan-Müller became an uncompromising moralist; *Adam Homo* (1841–48), a poetic epic, was a bitter contemporary satire. Hans Christian Andersen was most important for his fairy tales, the majority of whose plots were his own invention, though he also wrote novels, plays, travel books, and poems.

Kierkegaard

Søren Kierkegaard holds a position entirely isolated in Danish literature. His highly personal religious philosophy was expressed in such works as *Enten Eller* (1843; *Either/Or: A Fragment of Life*) and *Stadier paa livets vei* (1845; *Stages on Life's Way*). He spent his last years in a violent and passionate attack on "official Christianity."

Meïr Aron Goldschmidt edited a rebellious, anti-royalist weekly, *Corsaren* ("The Corsair"), while many of his novels and short stories were concerned with Jewish life in the Danish community. The 1850s and 1860s produced few new Danish writers of importance: the most original was Hans Egede Schack, whose novel *Phantasterne* (1857; "The Daydreamers") revealed great psychological gifts.

FAEROESE LITERATURE

Emergence of a written literature

Modern Faeroese literature emerged during the second half of the 19th century. Until this time, the literary tradition of the Faeroese was almost exclusively oral. It consisted principally of ballads, epic and fantastic in style and centred on legends such as that of Sigurd. A new, national written literature in Faeroese became possible only after the orthography was normalized by means of rules introduced in 1846 by the linguist and folklorist Venceslaus Ulricus Hammershaimb. Its development was promoted by nationalist agitation, which hastened the restoration of the old Faeroese parliament in 1852 and the end of the Danish royal trade monopoly in 1856.

Much of the writings of these early formative years consisted of patriotic poetry. The most memorable examples of such emotional songs were produced by Fríðrikur Petersen, Rasmus Effersøe, and Jóannes Patursson.

ICELANDIC LITERATURE

Literary and linguistic renaissance

The literary and linguistic renaissance in Iceland at the start of the 19th century was fostered by three men in particular: a philologist, Hallgrímur Scheving; a poet and lexicographer, Sveinbjörn Egilsson; and a philosopher and mathematician, Björn Gunnlaugsson. The principal movement in this renaissance was Romanticism. Inspired by the German philosopher Henrik Steffens, Bjarni Thorarensen produced nationalistic poetry that became a model for 19th-century lyrical poetry. Jónas Hallgrímsson, however, surpassed Thorarensen as a metrist. He was one of four involved in the periodical *Fjölnir* ("The Many-Sided"), which aimed to revolutionize literary theory and practice. The Fjölnismenn were anti-traditional and rejected the use of rhymes.

The group was replaced in the 1840s by another group of poets, of whom the most outstanding were Benedikt Gröndal, Steingrímur Thorsteinsson, and Matthías Jochumsson. Gröndal wrote powerful lyric poetry, two prose fantasies, and an autobiography, *Daegradvöl* (1923; "Day-Spending"). Thorsteinsson wrote nature poetry and satirical epigrams but is best remembered as translator of *King Lear* (1878) and *A Thousand and One Nights* (1857–64). Jochumsson's *Hallgrímur Péturssón* (1874) and hymn *Fadir andanna* (*c.* 1884; "Father of Spirits") established him as the greatest lyric poet of the three. He, too, translated Shakespeare in addition to Ibsen's *Brand.* A poet, Grímur Thomsen, was contemporary with but distinct from this group; his poetry was less lyrical, more austere and rugged, as *Hemings flokkur Áslákssonar* (1885; "The Story of Heming Aslakssonar") exemplifies.

The latter part of the century produced three talented poets: Thorsteinn Erlingsson, author of *Aldaslagur* (1911; "Sound of the Ages"); Einar Benediktsson, a Neoromantic; and Stephan G. Stephansson, an embittered expatriate whose irony passed in Iceland for Realism.

The 19th century also saw a renaissance in imaginative prose. Jón Thoroddsen wrote two novels that have acquired a position not incommensurate with that of the medieval sagas: *Piltur og stúlka* (1850; *Lad and Lass*) and the incomplete *Madur og kona* (1876; "Man and Woman"), distinguished in prose style, narrative skill, wit, and perceptive observation of peasant and small-town life.

The 20th century

NORWEGIAN

In the 1890s established Norwegian writers came under fire from the new generation. The manifesto of new ideas was an essay published in 1890 in the periodical *Samtiden* ("The Present Age") by Knut Hamsun, "Fra det ubevidste Sjaeleliv" ("From the Unconscious Life of the Mind"), which demanded attention to what was individual and idiosyncratic rather than typical. Hamsun was impatient with contemporary emphasis on social problems, and his early novels—*Sult* (1890; *Hunger*), *Mysterier* (1892; *Mysteries*), and *Pan* (1894)—exemplified these ideas; his later novels, such as *Markens grøde* (1917; *Growth of the Soil*), were less extreme but still showed a strong, sometimes savage irony. Hamsun won the Nobel Prize for Literature in 1920.

Knut Hamsun's emphasis on the individual

Lyric poetry at this time flourished with Sigbjørn Obstfelder, who had a close affinity with the Symbolist movement, and Nils Collett Vogt, who produced some of the best lyrics of the 1890s. In drama Gunnar Heiberg, who combined a sharply satirical wit with a lyric deftness, expressed the new spirit in *Kong Midas* (1890), *Gerts have* (1894; "Gert's Garden"), *Balkonen* (1894; "The Balcony"), and *Kjaerlighetens tragedie* (1904; "The Tragedy of Love"). Sharing Hamsun's preoccupation with the irrational side of human conduct was Hans E. Kinck, a writer of considerable power and penetration. In his verse drama *Driftekaren* (1908; "The Drover") and long novel *Sneskavlen brast* (1918–19; "The Avalanche Broke"), Kinck showed himself to be a more reflective and analytical writer than Hamsun.

The real achievements of Norwegian literature in the first half of the 20th century were in the novel and lyric poetry. Drama was not conspicuous, except for the plays of Gunnar Heiberg and Nordahl Grieg. In the early decades of the century, regionalism was a strong element, particularly in the novel; and authors adopted language coloured by dialect, thus becoming identified with their

region. Kristofer Uppdal, of the mid-north region of Trøndelag, wrote a remarkable work—a 10-volume novel cycle, *Dansen gjenom skuggeheimen* (1911–24; "The Dance Through the Shadow World"). The novel also treated of conflicts arising from the spread of industrialism, which Norway underwent later than did other European countries. The most proletarian writer was Oskar Braaten, but superior as an artist was Johan Falkberget, who wrote with understanding and historical insight about the miners in Røros in *Christianus Sextus* (1927–35) and in *Nattens brød* (1940; "Bread of Night"). Sigrid Undset, who won the Nobel Prize for Literature in 1928, set her novels in many different ages, and their concern was to examine women's loyalties within the framework of their role in society. A long historical novel, *Kristin Lavransdatter* (1920–22), was a masterpiece of Norwegian literature. Her later novels, *Gymnadenia* (1929; *The Wild Orchid*) and *Den braendende busk* (1930; *The Burning Bush*), were greatly influenced by her conversion to Roman Catholicism. Olav Duun, again of the mid-north region, revealed his insight into life as endless conflict in a six-volume novel cycle about the development of a peasant family through four generations—*Juvikfolke* (1918–23; *The People of Juvik*).

Central theme of Sigrid Undset's novels

Shortly before World War I, there were several good lyric poets: Herman Wildenvey, Olaf Bull, Tore Ørjasaeter, and Olav Aukrust. Between World Wars I and II, there emerged many socially committed writers: the poet Arnulf Øverland; a novelist and critic, Sigurd Hoel; a dramatist and critic, Helge Krog; and Nordahl Grieg. After World War II, Tarjei Vesaas wrote a remarkable series of novels, including the symbolic *Huset i mørkret* (1945; "The House in the Darkness") and *Bruene* (1966; "The Bridges"). Cora Sandel, who had made a major contribution with her "Alberte" trilogy (1926–39), continued to write, as did Aksel Sandemose, an experimental writer, and Johan Borgen, who won acclaim for his early short stories, the *Lillelord* trilogy (1955–57), and the autobiographical *Barndommens rike* (1965; "Childhood's Realm"). Borgen later became the leading novelist in Norway and maintained this standing until his death in 1979. Since then, Terje Stigen, Knut Faldbakken, and Bjørg Vik have become the dominant figures in prose fiction. Stigen's works are basically realistic narratives that variously treat historical and contemporary subjects. Faldbakken has demonstrated much fantasy and ingenuity, most recently having completed a series of novels that portray the collapse of technological society. An excellent short-story writer, Vik centres her attention on middle-class family life and often portrays it from a mildly feminist viewpoint.

SWEDISH

Mood of decadence and pessimism

The early years of the 20th century were a period of decadence and pessimism in Swedish literature. Representative of this mood were Hjalmar Söderberg and Bo Bergman. Söderberg's forte was the short story (*Historietter* [1898]), in which psychological subtlety and irony were happily combined and in which, as in his novels *Martin Bircks ungdom* (1901; "Martin Birck's Youth") and *Doktor Glas* (1905), he appeared as a master of Swedish prose. Bergman also produced memorable short stories, but his real medium was the lyric; he developed his talent in a series of collections from *Marionetterna* (1903; "The Marionettes") to *Riket* (1944; "The Kingdom").

The modern Swedish novel. The development of the novel was associated with Gustaf Hellström, Ludvig Nordström, Elin Wägner, and Sigfrid Siwertz. Hellström's work as a journalist in Europe, the United States, and England greatly influenced him. Irony and careful detail emerged in his best known novel, *Snörmakare Lekholm får en idé* (1927; *Lacemaker Lekholm Has an Idea*). Siwertz was a more elegant stylist, and a decisive influence upon him was the philosophy of Henri Bergson, reflected in *En flanör* (1914; "An Idler"); but his weightiest work was a family saga, *Selambs* (1920; *Downstream*), a novel of Stockholm during World War I. Nordström, overflowing with vitality and keen but grotesque humour, accomplished some of his best work in *Landsorts-bohème* (1911; "Small-Town Bohemia") and in his short stories—*e.g.*, *Fiskare* (1907; "Fishermen") and *Öbacka-bor* (1921). Elin Wägner was an ardent pacifist and feminist; her most powerful work was a peasant novel, *Åsa-Hanna* (1918). The outstanding novelist of the 1920s was Hjalmar Bergman: with vivid imagination and restless energy, Bergman wrote a long series of stories, many set in "Wadköping" (his native Örebro), others in Italy. In *Loewenhistorier* (1913) he depicted an irrational, impulsive, unsuccessful hero; in *Farmor och vår Herre* (1921; *Thy Rod and Thy Staff*) he portrayed one of the dominating female personalities that fascinated him. The satire *Markurells i Wadköping* (1919; *God's Orchid*) and *Swedenhielms* (performed 1925), one of the few Swedish comedies, were his most widely known works.

The "proletarian" novel

Meanwhile, the "proletarian" novel had been developed by writers concerned with the miseries of the working class, particularly Martin Koch and Ivar Lo-Johansson. There was particularly harsh criticism of working class conditions in stories by Jan Fridegård. Vilhelm Moberg wrote novels of peasant life but achieved his greatest success with the four-part prose epic about a group of Swedish emigrants to North America, *Utvandrarna* (1949–59; *The Emigrants*). The development of the Swedish autobiographical novel was helped by Eyvind Johnson, with the series "Romanen om Olof" (1934–37); Harry Martinson, with *Nässlorna blomma* (1935; *Flowering Nettle*) and *Vägen ut* (1936; "The Way Out"); and Agnes von Krusenstjerna. In her novel cycles, the "Tony" trilogy (1922–26) and the "Fröknarna von Pahlen" series (1930–35), Krusenstjerna described her own aristocratic environment and analyzed a degenerate psychology. Harry Martinson was one of a group of five primitivist writers formed about 1930. He later developed into one of the finest lyricists of the century. Sensuous imagery and a feeling for nature characterized his work. He attempted to revive the verse epic in his *Aniara* (1956), a symbolical story of a voyage of a spaceship.

Pär Lagerkvist's work

The internationally best known Swedish writer of the 20th century was Pär Lagerkvist, who won the Nobel Prize for Literature in 1951. In his youth a bold innovator, he later developed an admirably pure prose style, as in the allegorical novel *Dvärgen* (1944; *The Dwarf*). His collections of poems, *Ångest* (1916; "Anguish"), and early plays—for instance, *Himlens hemlighet* (1919; "The Secret of Heaven")—were Expressionistic in style. The dominant theme throughout Lagerkvist's work was a search for vital, often outspokenly religious values.

Development of lyric poetry. Several of the best Swedish writers were connected with the development of lyric poetry. One of the most notable, Vilhelm Ekelund, was in his youth the chief exponent of Symbolism in Sweden and later, as an author of aphorisms, exerted much influence on the development of literary modernism. Among the most popular poets were Dan Andersson, Birger Sjöberg, and Hjalmar Gullberg. In Gullberg's poetry, religious commitment and classical learning are balanced by irony and wit. A more esoteric style in modernism was introduced by Bertil Malmberg and developed by the group of poets called the generation of the 1940s, which included Erik Lindegren and Karl Vennberg. Stylistically influenced by T.S. Eliot, they often expressed an anguish and disbelief that approached French Existentialism. Lindegren's *Mannen utan väg* (1942; *The Man Without a Way*) was typical of this generation's search for meaning in life. The most distinguished novelist of the 1940s was Lars Ahlin, who was concerned with man's search for grace through love and humiliation in works such as *Min död är min* (1945; "My Death Is Mine").

The greatest lyric poet of the century was Gunnar Ekelöf. His first collection of poems, *Sent på jorden* (1932; "Late on Earth"), was heralded as the first specimen of Surrealism in Swedish literature. Ekelöf's later development passed through successive phases of Romanticism and anti-poetic Skepticism resolved in a trilogy of books blending autobiography and Eastern mysticism.

Contemporary trends. In reaction to the literature of the 1940s and 1950s, which was much concerned with artistic form and the individual approach to life, the 1960s was a period of political and social commitment in poetry and fiction alike. Recurrent topics were the war in Vietnam and bitter onslaughts on the Swedish welfare state. Inde-

Emphasis on political and social issues

pendent lyric poetry, however, continued to be produced by writers such as Östen Sjöstrand and Thomas Tranströmer, and a tortured experience of life, coloured by Roman Catholicism, was forcefully expressed in the novels of Birgitta Trotzig.

The aforementioned movement toward and subsequently away from profoundly politically committed literature is exemplified by the work of Sara Lidman. During the 1950s she was one of Sweden's most creative novelists but then ceased producing fiction in order to take part in the political debate of the time. Since the late 1970s, however, Lidman has returned to creative literature with a series of novels centred on life in an isolated Swedish community. Sven Lindqvist went through a similar process; after a period of committed writing, he returned in *En älskares dagbok* (1981; "A Lover's Diary") to a more or less autobiographical novel of his own youth. Political writing persists in Sweden, but it has become more imaginative and less tied to immediate events. P.C. Jersild, for example, has painted a chilling picture of civilization after a devastating nuclear war in *Efter floden* (1982; "After the Flood"); he had earlier demonstrated his talent in allegories set in a state veterinary institution and in a hospital. Sven Delblanc also has made use of allegory, and there is sometimes an almost mystical intensity apparent in his work.

Developments in Finno-Swedish literature. The second flourishing of Finno-Swedish literature occurred in the 1920s, with the development of modernism in lyric poetry. This trend was initiated by Edith Södergran, whose visionary, dreamlike poems proved influential throughout much of Scandinavia. After her came such poets as Gunnar Björling, noted for his impressionistic pictures of nature; Rabbe Enckell, a key theoretician of the movement; and Elmer Diktonius, devoted to political concerns.

Among the most talented prose writers of the period was Runar Schildt, whose short stories dealt with such questions as the relation of the intellectual and the artist to life. Several outstanding writers appeared somewhat later in the century, as, for example, Tito Colliander, who treated themes of guilt and atonement from a religious standpoint; Christer Kihlman, whose novels combined social criticism with psychological commentary; and Tove Jansson, internationally famous for her imaginative portrayals of the fairy-tale realm of Moomintrolls.

DANISH

The influence of Georg Brandes. About 1870 there arose in Denmark a new movement, led by Georg Brandes, from which a modern (*i.e.*, a Naturalistic or Realistic) literature emerged. His *Hovedstrømninger i det 19de aarhundredes litteratur* (1872–90; *Main Currents in 19th Century Literature*), describing the growth and defeat of reaction, caused a great sensation. As noted earlier, he influenced Ibsen and Strindberg and wrote many scholarly and critical works illustrating radical ideas. His later biographies of Shakespeare, Goethe, Voltaire, Julius Caesar, and Michelangelo revealed how he was influenced by Nietzsche into developing a philosophy of aristocratic radicalism. Among his followers were Jens Peter Jacobsen, whose short story "Mogens" (1872) and novel *Fru Marie Grubbe* (1876) are the supreme examples of Danish Naturalism, while his other novel *Niels Lyhne* (1880) and some of his short stories dealt with dream as against reality; and Holger Drachmann, greatest lyric poet of the period, who later reacted strongly against Brandes and whose poetry and prose were often about the sea.

Georg Brandes' radical ideas

Henrik Pontoppidan, one of Denmark's greatest novelists, dealt at first with social injustices and contemporary political, moral, and religious problems in his short stories. The Denmark of his day was also the subject of his greatest work, three long novel cycles, *Det forjaettede land* (1891–95; *The Promised Land*), *Lykke-Per* (1898–1904; "Lucky Peter"), and *De dødes rige* (1912–16; "The Realm of the Dead"); and in these he makes penetrating, if unflattering, analyses of Danish national character. Herman Bang was another novelist interested in the outsiders of life and in insignificant people. His skillful, mainly impressionistic technique was displayed in his best novels, *Ved vejen* (1886; "By the Way-Side"), *Tine* (1889), and *Det hvide hus* (1898; "The White House").

Other notable writers at the end of the century were Gustav Wied, whose "satyr plays" and whose novels *Livsens Ondskab* (1899; "Life's Malice") and *Knagsted* (1902) were full of malicious humour; Vilhelm Topsøe, a conservative realist; Peter Nansen, who wrote stories reminiscent of those of Guy de Maupassant; Carl Ewald, whose nature stories were based on Darwinian philosophy; Karl Larsen, who caught the atmosphere of Copenhagen and its inhabitants with fine precision; and several playwrights, including Edvard Brandes, Otto Benzon, Gustav Esmann, Sven Lange, Einar Christiansen, and Henri Nathansen.

Neoromantic revival. In the 1890s a Neoromantic poetic revival occurred, reinstating the value of emotion and fantasy. The leader of these Symbolist poets was Johannes Jørgensen, whose finest works show a simplicity of style and intensity of feeling. Other poets of the time included Viggo Stuckenberg, who expressed sad resignation; Sophus Claussen, whose poems, often obscure, show sensuality, pantheistic love of nature, and sophisticated aestheticism; and Helge Rode, a mystic who also wrote plays and criticism attacking intellectualism.

Symbolist mode

20th-century literary trends. Several women contributed to literature at the turn of the century: Gyrithe Lemche, who wrote a novel cycle, *Edwardsgave* (1900–12); Agnes Henningsen, a brilliant writer who was often concerned with experiences of the emancipated woman; and Karin Michaëlis, a fine psychologist, best known for her novel *Den farlige alder* (1910; *The Dangerous Age*).

The two greatest early 20th-century novelists were Martin Andersen Nexø and Johannes Vilhelm Jensen. Nexø's works described the lives of poor people; *Pelle Erobreren* (4 vol., 1906–10; *Pelle the Conqueror*) and *Ditte Menneskebarn* (3 vol., 1917–21; *Ditte: Daughter of Man*) were great epics of proletarian life, and his reminiscences were among the finest in the language. Jensen, who was also a great and original lyric poet and prolific essayist, wrote *Den lange rejse* (6 vol., 1908–22; *The Long Journey*), an ambitious epic of man from the baboon stage to the discovery of America; he was also noted for *Himmerlandshistorier* (1904, revised 1910; "Tales from Himmerland"), based on his childhood memories of North Jutland; *Kongens fald* (1900–01; *The Fall of the King*); and nine volumes of *Myter* (1907–44; "Myths"). Other novelists of this period included Jakob Knudsen, whose interest, taking account of the inequality of man and the need for authority, was with Christian and moral problems; Harald Kidde, an introspective and melancholy writer; and Knud Hjortø, a keen and intelligent writer of psychological novels.

Nexø's epics of proletarian life

Regional literature of the early 1900s was produced chiefly by Jutland writers. Prominent among them were three poets: Jeppe Aakjaer, Johan Skjoldborg, and Thøger Larsen. Also, Marie Bregendahl and Harry Søiberg drew upon Jutland settings for their novels.

Significant poets of the post-World War I generation were Tom Kristensen, Otto Gelsted, Emil Bønnelycke, Kai Friis Møller, and Per Lange. Among interesting novelists, Jacob Paludan wrote some very good fiction—*Fugle omkring fyret* (1925; *Birds Around the Light*) and *Jørgen Stein* (1932–33)—as also did Hans Kirk, whose *Fiskerne* (1928; "The Fishermen") was social realism at its best. Harald Herdal, a disciple of Nexø, exposed society's hypocrisy in his proletarian novels. Jørgen Nielsen's themes were suppressed hatred, sin, and fear among Jutland peasants. H.C. Branner, an important writer of novels, plays, and short stories, spoke of the loneliness of men and the danger of power. Another writer of these three genres was Knud Sønderby, who had a brilliant style and deep understanding. Nis Petersen, poet and novelist, was famous for *Sandalmagernes gade* (1931; *The Street of the Sandalmakers*) and *Spildt maelk* (1934; *Spilt Milk*). Isak Dinesen (Karen Christence Dinesen, Baroness Blixen-Finecke), an aristocratic writer with subtle irony and unusual sensitivity, wrote both in Danish and in English. Her first notable work, a collection of short stories featuring a strong fairy-tale-like quality, was in fact written in English. Entitled *Seven Gothic Tales*, it was published in 1934 in the United States and subsequently translated

Isak Dinesen

by the author into Danish as *Syv fantastiske Fortællinger*. Other major works of Dinesen include her memoir *Den afrikanske Farm* (1937; *Out of Africa*) and two more collections of finely crafted stories, *Vinter-Eventyr* (1942; *Winter's Tales*) and *Sidste Fortællinger* (1957; *Last Tales*).

Two Faeroese novelists also made important contributions to modern Danish prose fiction: Jørgen-Frantz Jacobsen, with his novel *Barbara* (1939), which provides a fascinating portrait of a capricious woman; and William Heinesen, with his masterpiece *De fortabte spillemænd* (1950; *The Lost Musicians*). Here, as in the rest of his varied writings, Heinesen renders Faeroese life as a microcosm illustrative of social, psychological, and cosmic themes. Other distinguished novelists of the time were Hans Scherfig, a great humorist and social satirist; and Martin Alfred Hansen, a psychological novelist, whose best known novel is *Løgneren* (1950; *The Liar*).

Danish playwrights of the post-World War I period such as Sven Clausen and Svend Borberg were influenced by German Expressionism, Symbolism, Luigi Pirandello, and Sigmund Freud. Kaj Munk, who revived the heroic drama of Shakespeare and Schiller, showed unusual qualities in his best plays—*En idealist* (1928; *Herod the King*) and *Ordet* (1932; *The Word*)—which were concerned with problems of God and man. The work of Kjeld Abell marked a severance from Naturalist drama, and a radical perspective underlay his witty dialogue. His most important plays were *Melodien, der blev vaek* (1935; *The Melody That Got Lost*), *Anna Sophie Hedvig* (1939), *Dage på en sky* (1947; *Days on a Cloud*), and *Skriget* (1961; "The Scream"). C.E. Soya was an important playwright of the period, and also a novelist and fine short-story writer; some of his daring experiments with the theatre have been very successful.

The modern Danish theatre

Many postwar poets found an aesthetic manifesto in *Fragmenter af en dagbog* (1948; "Fragments of a Diary"), by Paul la Cour, who was influenced by contemporary French poetry. Jens August Schade, a sophisticated naivist, also had an important influence, as did the prematurely deceased poets Gustaf Munch-Petersen and Morten Nielsen. A revival of poetry followed the liberation (1945), and the Existentialist periodical *Heretica* (1948–53) became the voice of a group of young writers who regarded a Christian philosopher, Vilhelm Grønbech, as their spiritual progenitor. Two outstanding poets apart from the *Heretica* group were Halfdan Rasmussen, who also wrote excellent nonsense verse, and Erik Knudsen, also a brilliant satirical playwright. Both studied contemporary problems and reacted against the anti-rationalism and anti-intellectualism of the *Heretica* movement. Tove Ditlevsen was another important poet, as well as novelist and short-story writer, unattached to any group; his often intensely personal work reflects the loneliness of life in the poorer quarters of Copenhagen. Klaus Rifbjerg has been the dominant novelist in Denmark since the publication of his *Den kroniske uskyld* ("Chronic Innocence") in 1958. He is a writer of great inventiveness and linguistic originality, who has analyzed modern society and its problems, both public and private, in realistic novels. In a more satirical vein, Leif Panduro has examined the place of the individual in society, paying special attention to the problems of middle age and the emptiness of a welfare-state society in novels and television dramas. A more philosophical approach is found in Villy Sørensen, whose Kafkaesque stories place him in the sphere of Absurd literature. There has been, at the same time, a vogue of the documentary novel, the principal exponent of which has been Thorkild Hansen.

The modern poetry that became a hallmark of Danish literature after World War II gave rise to the mature works of Frank Jæger, Thorkild Bjørnvig, and Ivan Malinovski. Further experimentation resulted in Structuralist works such as Inger Christensen's *Det* (1968; "It"). Henrik Nordbrandt has been the leading Danish poet of the 1970s and '80s. A pessimistic sense of isolation pervades his poems, which have been influenced by Oriental writings.

The 1970s saw an upsurge of women authors, some consciously writing for the feminist cause and others content to portray life as experienced by women. Outstanding among them are Dea Mørch, Kirsten Thorup, and Dorit Willumsen.

FAEROESE

Faeroese literature came into its own after the turn of the century. Jens H.O. Djurhuus, who created rhetorical poetry of splendid resonance, was the first to emerge as a writer of international stature. His brother, Hans Andrias Djurhuus, wrote in a more naive manner, producing poems, fairy tales, and plays, which were based on native historical traditions and legends.

Five writers dominated the Faeroese literary scene from about the 1930s through mid-century. Of these so-called Faeroese golden age authors two, Jørgen-Frantz Jacobsen and William Heinesen, wrote in Danish (see above), while the other three, Christian Matras, Heðin Brú (Hans Jakob Jacobsen), and Martin Joensen, in Faeroese. The works of Matras reveal a profound lyric poet seeking to interpret the essence of Faeroese culture. A fine stylist, Brú did much to create a Faeroese literary prose in his portrayals of village life in a time of transition (*e.g., Feðgar á ferð* [1940; *The Old Man and His Sons*]). Joensen's novels and short stories are of a similar character, but their emphasis is on psychological realism rather than on style. A prose writer of a distinctly more modern bent is Jens Pauli Heinesen. His works reflect an approach to Faeroese life that is generally more international than that of Brú or Joensen and that is infused with a certain satirical element.

Faeroese golden age writers

Poetry continues to attract many writers. Karsten Hoydal was the first Faeroese writer to compose verse directly influenced by modern foreign poets; he also translated many of their works, especially those of Edgar Lee Masters of the United States. Regin Dal and Steinbjørn B. Jacobsen have gone much further in their modernism, the latter adopting a style somewhat akin to the Imagism of the U.S. poet Ezra Pound. Other writers whose works exhibit a modernist tendency include Guðrið Helmsdal, the foremost Faeroese woman poet of the contemporary scene.

ICELANDIC

Modern Icelandic prose writing did not really develop until the late 1870s, when a group of young men, influenced by the theories of the Danish critic Georg Brandes, began their literary careers. Unfortunately, they had absorbed Brandes' ideas uncritically, which resulted in introspective, self-pitying works believed by their authors to be realistically written. The early works of Einar Kvaran suffered from this flaw, but Kvaran later developed into a novelist of skill and power.

Notable 20th-century prose writers. Several writers of this time showed a keen eye for character and an understanding of human feelings and of the stark life of rural Iceland: Jón Trausti (Gudmundur Magnússon), who wrote the cycle *Heidarbýlid* (4 vol., 1908–11; "The Mountain Cot"); Gunnar Gunnarsson, whose *Kirken på bjerget* (1923–28; "The Church on the Mountain") was written in Danish; and Gudmundur Hagalín. The outstanding modern prose writer was Halldór Laxness, who was awarded the Nobel Prize for Literature in 1955. His mature works were influenced by his conversion to Roman Catholicism and his identification with the basic ideas of Communism. His major works were *Salka Valka* (1936), *Sjálfstaet fólk* (1935; *Independent People*), *Íslandsklukkan* (1943; "The Bell of Iceland"), and *Gerpla* (1952). He helped restore Icelandic as a sensitive medium for storytelling.

The works of Halldór Laxness

Among more recent prose fiction writers, Guðbergur Bergsson has proved himself one of the most talented and forceful. Reflective of the growing social and political consciousness of the 1960s, some of his novels from that period—*Ástir samlyndra hjóna* (1967; "The Love of a Harmoniously Married Couple") and *Anna* (1969)—subjected contemporary Icelandic society and the military relations of the nation with the United States to biting satirical attacks. His later works, the collection of short stories *Hvað ereldi guðs* (1970; "What Does God Eat") and a series of novels produced in the mid-1970s, were decidedly experimental in character, revealing an attempt by the author to go beyond ordinary reality to expose some of the more disgusting and grotesque aspects of life.

Major poets. At the beginning of the 20th century, poetry had lyricists in Thorsteinn Erlingsson, whose early delicacy later developed into a more powerful note in

"Aldaslagur" (1911; "Sound of the Ages") and in an incomplete epic, "Eidurinn" (1913; "The Oath"); in Einar Benediktsson, who wrote in an ornate style sometimes capable of greatness, as "Í dísarhöll" ("In the Hall of the Muses") shows; and in Stephan G. Stephansson, an expatriate farmer in Canada who was a more bitter poet, influenced by the "realism" that passed for Georg Brandes' ideas in Icelandic literature—but *Andvökur* (1909–38; "Sleepless Nights") revealed a sensitive spirit.

Prominent poets of the next generation included Davíd Stefánsson, a traditionalist who expressed deep personal feelings in straightforward language and simple verse forms. His approach was shared by Tómas Gudmundsson and Jón Helgason, whose book *Úr landsudri* (1939, revised 1948; "From the South") was outstanding. Steinn Steinarr (Adalsteinn Kristmundsson), who was deeply influenced by Surrealism, experimented with abstract styles and spearheaded modernism in Icelandic poetry with his collection *Ljóð* (1937; "Poems").

Since mid-century several poets have distinguished themselves. The early works of Hannes Pétursson showed great sensitivity and skill in adapting Icelandic to new, European metres. Pétursson's more recent poems (those in the collection *Ur hugskoti* [1976; "Recollections"]), however, reveal a movement away from innovative forms to more traditional verse. Still other contemporary poets of merit include Thorsteinn frá Hamri and Sigurður Pálsson. Hamri's poems *Veðrahjálmur* (1972; "Sun Rings") grapple with questions about lasting values, particularly with the possibility of realizing human fellowship in the modern world. Pálsson's *Ljóð vega salt* (1975; "Poems on the See-Saw") combine autobiographical elements with philosophical questioning about the nature of contemporary life.

The development of the Icelandic drama. Icelandic drama really started to develop through Jóhann Sigurjónsson, whose first success was *Fjalla-Eyvindur* (1911; *Eyvind of the Hills*), followed by *Galdra-Loftur* (1915; "Loftur the Sorcerer"); both plays were based on powerful folktales. Gudmundur Kamban's *Hadda-Padda* (1914) was highly praised by Georg Brandes, and he remained important in Scandinavian drama for the next quarter of a century. After Kamban, there were few plays of lasting value, though Davíd Stefánsson's *Gullna hlidid* (1941; "The Golden Gate"), Jakob Jónsson's *Tyrkja-Gudda* (published 1948), and Agnar Thórdarson's satiric comedy of modern Reykjavík life, *Kjarnorka og kvenhylli* (1957; "Nuclear Force and Female Popularity"), had considerable merit. In *Gankstklukkan* (1962; *The Cuckoo Clock*) the latter produced a powerful play on the dehumanizing effect of modern life.

BIBLIOGRAPHY. Broad coverage of Scandinavian literature is found in ELIAS BREDSDORFF, BRITA MORTENSEN, and RONALD POPPERWELL, *An Introduction to Scandinavian Literature, from the Earliest Time to Our Day* (1951, reissued 1970), offering a useful short survey; FREDERICK J. MARKER and LISE-LONE MARKER, *The Scandinavian Theatre: A Short History* (1975); and VIRPI ZUCK (ed.), *Dictionary of Scandinavian Literature* (1990), assembling author entries, topical articles, and an extensive bibliography subdivided by language and category.

The literature of individual nations is treated in the following surveys: P.M. MITCHELL, *A History of Danish Literature*, 2nd augmented ed. (1971); SVEN H. ROSSEL (ed.), *A History of Danish Literature* (1992); THOMAS WARBURTON, *Åttio år finlandssvensk litteratur* (1984); STEFÁN EINARSSON, *A History of Icelandic Literature* (1957); HARALD BEYER, *A History of Norwegian Literature* (1956, reissued 1979; originally published in Norwegian, 1952); HARALD S. NAESS (ed.), *A History of Norwegian Literature* (1993); and ALRIK GUSTAFSON, *A History of Swedish Literature* (1961), an excellent critical history, with bibliographic appendix.

Early Scandinavian literary history is the focus of GABRIEL TURVILLE-PETRE, *The Heroic Age of Scandinavia* (1951, reprinted 1976), and *Origins of Icelandic Literature* (1953, reissued 1975); CAROL J. CLOVER, *The Medieval Saga* (1982); CAROL J. CLOVER and JOHN LINDOW (eds.), *Old Norse–Icelandic Literature: A Critical Guide* (1985); JÓNAS KRISTJÁNSSON, *Eddas and Sagas: Iceland's Medieval Literature*, 2nd ed. (1992); S.B.F. JANSSON, *The Runes of Sweden* (1962); ANTON BLANCK, *Den nordiska renässansen i sjuttonhundratallets litteratur* (1911); JAMES A. PARENTE, JR., and RICHARD ERICH SCHADE (eds.), *Studies in German and Scandinavian Literature After 1500* (1993); REINHOLD AHLÉEN, *Swedish Poets of the Seventeenth Century* (1932); and ALBERT NILSSON, *Svensk romantik: den platonska strömningen* (1916).

Noteworthy studies covering more recent years include RICHARD BECK, *History of Icelandic Poets, 1800–1940* (1950, reprinted 1966); STEFÁN EINARSSON, *History of Icelandic Prose Writers, 1800–1940* (1948, reprinted 1966); JANET GARTON, *Norwegian Women's Writing, 1850–1990* (1993); JAMES WALTER MCFARLANE, *Ibsen and the Temper of Norwegian Literature* (1960, reprinted 1979); HELGE G. TOPSÖE-JENSEN, *Scandinavian Literature from Brandes to Our Day* (1929, reprinted 1971; originally published in Danish, 1928); BRIAN W. DOWNS, *Modern Norwegian Literature, 1860–1918* (1966), an excellent survey with bibliography; SVEN H. ROSSEL, *A History of Scandinavian Literature: 1870–1980* (1982; originally published in German, 1973); BODIL WAMBERG (ed.), *Out of Denmark: Isak Dinesen/Karen Blixen, 1885–1985, and Danish Women Writers Today*, trans. from Danish (1985); KARIN ELKJAER and POUL ZERLANG (eds.), *Danske og udenlandske forfattere efter 1914*, 4th ed. (1977); MARTIN S. ALLWOOD (ed.), *20th Century Scandinavian Poetry* (1950); IRENE SCOBBIE (ed.), *Aspects of Modern Swedish Literature* (1988); SARAH DEATH and HELENA FORSÅS-SCOTT (eds.), *A Century of Swedish Narrative* (1994); and FAITH INGWERSEN and MARY KAY NORSENG (eds.), *Fin(s) de siècle in Scandinavian Perspective* (1993).

JOHN M. WEINSTOCK and ROBERT T. ROVINSKY (eds.), *The Hero in Scandinavian Literature: From Peer Gynt to the Present* (1975), treats film along with literary genres. JANET MAWBY, *Writers and Politics in Modern Scandinavia* (1978), emphasizes the impact of the German occupation and the U.S. involvement in Vietnam. JOHN L. GREENWAY, *The Golden Horns: Mythic Imagination and the Nordic Past* (1977), studies myth and its effect on Scandinavian literature and life. JESSE L. BYOCK, *Feud in the Icelandic Saga* (1982), and *Medieval Iceland* (1988), uses Icelandic sagas as sources for analysis of social and economic history.

Recommended anthologies are ELIAS BREDSDORFF (ed.), *Contemporary Danish Plays* (1955, reissued 1970), and *Contemporary Danish Prose* (1958, reprinted 1974); *Modern Nordic Plays: Denmark* (1974); HEDIN BRØNNER (trans. and ed.), *Faroese Short Stories*, trans. from Faroese and Danish (1972); JANET GARTON and HENNING SEHMSDORF (trans. and eds.), *New Norwegian Plays* (1989); FREDERIC FLEISHER (trans. and ed.), *Seven Swedish Poets* (1963); and GUNILLA M. ANDERMAN (compiler), *New Swedish Plays* (1992).

(B.S.B./S.Be./J.W.McF./B.Mo./E.O.G.T.-P./W.G.J.)

Classical Scholarship

Classical scholarship comprises the study, in all its aspects, of ancient Greece and Rome. In continental Europe this field is known as "classical philology"; the use, in some circles, of "philology" to denote the study of language and literature—the result of abbreviating the 19th-century "comparative philology"—has lent an unfortunate ambiguity to the term. During the 19th century, Germans evolved the concept of *Altertumswissenschaft* ("science of antiquity") to emphasize the unity of the various disciplines of which the study of the ancient world consists. Broadly speaking, the province of classical scholarship is in time the period between the 2nd millennium BC and AD 500 and in space the area covered by the conquests and spheres of influence of Greece and Rome at their widest extent.

This article surveys the history of classical scholarship thus defined from antiquity until the late 20th century. For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 911, 912, 10/11, 10/21, 10/31, 10/35, 10/42, 10/51, and 10/53, and the *Index*.

The article is divided into the following sections:

Antiquity and the Middle Ages 22
- Greek scholarship 22
- Latin scholarship 25

The revival of learning 26
- Renaissance humanism 26
- Beginnings of modern scholarship 26
- The Renaissance outside Italy 26
- Scholarship in the 17th century 27
- The 18th century: the age of Bentley 27

Modern classical scholarship 28
- The new German humanism 28
- The rise of textual criticism 28
- Developments in the study of ancient history and philosophy 29
- Developments in archaeology and art history 29
- The rise of professionalism 30
- Late 19th-century developments in German scholarship 30
- Classical scholarship in the 20th century 30

Bibliography 31

ANTIQUITY AND THE MIDDLE AGES

Until the Renaissance, Greek scholarship in the East and Latin scholarship in the West tended to follow different courses, and it is therefore convenient to treat them separately during this period.

Greek scholarship. *Beginnings.* Greek epic poetry was recited in early times by professional performers known as rhapsodists, or rhapsodes, who sometimes offered interpretations of the works as well. In the 6th century BC Theagenes of Rhegium is said to have "searched out Homer's poetry and life and date," to have offered an allegorical interpretation of the battle of the gods in the 20th book of the *Iliad*, and to have been cited for a variant in Homer's text. The Sophists of the 5th century BC—paid writers, lecturers, and teachers such as Protagoras, Prodicus, Gorgias, and Hippias—gave ethical instruction in the form of the exposition of poetry, particularly that of Homer, which from this time formed the staple of Greek education. Some of them were interested in etymology, phonetics, the exact meanings of words, correct diction, and the classification of the parts of speech. Hippias laid the foundations of ancient chronography by making a list of victors in the Olympic Games, and Alcidamas (*c.* 400 BC) wrote a book on Homer. However, the efforts of the Sophists in this direction, considerable as they were, had a more or less casual and arbitrary character.

Exposition of Homeric poetry

Plato (*c.* 428/427–348/347 BC) strongly resisted the claim that the poets were reliable interpreters of religion and morality. In his dialogue *Cratylus* he rejected the theory that the study of words can reveal the meaning of things, insisting that things themselves must be studied. Plato's pupil Aristotle (384–322 BC) defended poetry against his master; he valued highly the *Iliad* and the *Odyssey*, which from his time were regarded (together with the mock-epic *Margites*) as the genuine works of an individual Homer. He took a similar view of tragedy, which he believed effected a purification (*katharsis*) of the emotions upon which it played. Aristotle wrote about linguistic, dramatic, and other problems in Homer, refuting such detractors of the poet as Zoilus, compiled lists of Olympic and Pythian victors, collected details about the Athenian tragic and comic festivals, and supplemented his *Politics* with a collection of 158 studies of the constitutions of various Greek states. He also carried further the discussion of the constituent parts of a sentence and discussed the nature of synonyms, compounds, and rare words in early poetry.

The school of Aristotle

The school of Aristotle, known as the Lyceum, or Peripatos, continued to make this kind of learned work an adjunct to its philosophical activities. Aristotle's successor, Theophrastus (*c.* 372–*c.* 287 BC), collected the opinions of earlier philosophers. Dicaearchus (flourished *c.* 320 BC) wrote about the life of Greece, and Aristoxenus (flourished late 4th century BC) about the history and the theory of music. Heracleides Ponticus (*c.* 390–*c.* 322 BC) wrote one book on Archilochus and Homer and another on the dates of Homer and Hesiod. Clearchus collected proverbs, and Demetrius of Phaleron fables. All these philosophers were guided by Aristotle's teleological concept of intellectual activity, according to which philosophy is the culminating element of civilization. A 4th-century commentary on an Orphic poem, discovered in 1963 on a papyrus from a grave in Derveni, Macedonia, deserves mention as the earliest known commentary on a text; it is not a linguistic commentary but offers an allegorical interpretation that is doubtless very different from what the poet had intended.

The Library of Alexandria

During the Hellenistic Age (usually reckoned to extend from the death of Alexander the Great in 323 BC to the 1st century AD) scholarship flourished nowhere more than in the great city of Alexandria, the capital of the Ptolemies, the kings of Egypt. Early in the 3rd century BC Ptolemy I founded the famous Mouseion (Museum) of Alexandria, a community of learned men organized along the lines of a religious cult and headed by a priest of the Muses; part of the Museum was a splendid library that became the most celebrated of the ancient world. In its establishment the king is said to have had the assistance of the eminent Peripatetic scholar and statesman Demetrius of Phaleron, who left Athens about 300 BC; unfortunately, the evidence about the part he played is scanty and unreliable. The Museum community included both poets and scholars, as well as several individuals who combined these pursuits. From the time of the poet-scholar Philetas, or Philitas (*c.* 330–*c.* 270 BC), the tutor of Ptolemy II, the scholars there were much concerned with the collection and interpretation (*glossae*) of rare poetic words. Philetas' pupil Zenodotus of Ephesus (*c.* 325–260 BC) was the first librarian at Alexandria; using the manuscripts collected for the Library but also trusting to his own judgment, sometimes in a manner that seemed to later critics dangerously subjective, he made the first critical edition of Homer, marking passages of doubtful authenticity with critical signs in the margins. Zenodotus also edited Pindar

and Anacreon and perhaps other lyric poets; at about the same time the epic and elegiac poet Alexander Aetolus is said to have edited the tragic poets, and the dramatic poet Lycophron the comic poets, but singularly little is known about these editions.

Somewhat later the great poet Callimachus (*c.* 305–*c.* 240 BC) compiled the *Pinakes* ("Tablets"), a vast catalogue raisonné of the chief authors, with biographical and bibliographical information. Callimachus is said to have written a book opposing the chief Peripatetic critic of the time, Praxiphanes, and is widely held to have criticized Peripatetic literary theory; but the scantiness of the evidence for this enjoins great caution.

Rather later the great geographer and mathematician Eratosthenes (*c.* 276–*c.* 194 BC), the third librarian, laid the foundations of a systematic chronography; more of his work would be known had it not been largely superseded in popular use by the 2nd-century chronicles of Apollodorus of Athens, which were a learned compilation but left out the important scientific and mathematical part.

Zenodotus' editions of Homer and Hesiod were improved upon by the fourth librarian, Aristophanes of Byzantium (*c.* 257–180 BC), who also edited the lyric poets, setting out their verses according to a systematic metrical theory; edited Aristophanes, Menander, and perhaps other comic poets; edited Sophocles and at least part of Euripides; and compiled useful summaries of the plots of plays with details of their productions. His *Lexeis* ("Readings") was the most important of the numerous lexicographical works produced at this time, which included lexicons of particular authors and dialects; he also wrote some of the many treatises about literature that were now appearing.

Aristarchus of Samothrace (*c.* 217–145 BC), the sixth librarian, wrote not only monographs about poetry but also important commentaries on Homer, Pindar, and much of tragedy and comedy. Aristarchus was one of the many learned men who left Alexandria in consequence of the disastrous persecution of learning by Ptolemy VIII, from which that city's standing as a great centre of learning never quite recovered. (The great library survived a fire set in Alexandria in 47 BC by Julius Caesar, whose army supported Cleopatra in a civil war. It was finally destroyed in AD 391 by the patriarch Theophilus of Alexandria.)

During the 3rd century BC the Stoics, particularly Chrysippus (*c.* 280–*c.* 206 BC), made important contributions to the study of grammar, linked with the development of Stoic logic. Early in that century the Stoic Crates of Mallus emigrated to the court of King Eumenes II of Pergamum, which the Attalid dynasty had begun to make into a literary centre comparable with, though hardly equal to, Alexandria. Crates probably wrote commentaries on the *Iliad* and the *Odyssey,* characterized by the allegorical interpretation, faith in the accuracy of Homer's geography, and grammatical rigour typical of the Stoic school. Under Stoic influence the Pergamenes tended to stress the element of anomaly in grammar, while the Alexandrians stressed the element of analogy; that is, the Alexandrians insisted on the natural, inherent orderliness of grammar, while the Pergamenes approached the subject as empiricists, being content to organize observations of actual usage into a body of knowledge. But the details of the alleged controversy over this matter are obscure and known largely from suspiciously late sources. If the extant grammar ascribed to Dionysius Thrax, a pupil of Aristarchus active about 120 BC, is genuine, then the Alexandrian school of grammar was by that time already considerably influenced by the Stoics.

During the 1st century BC, by which time Rome was beginning to be the chief centre of Greek scholarship, Philoxenus wrote on Greek dialects, among which he included Latin; he was the first scholar to be aware of the existence of monosyllabic roots. Under Augustus, Tryphon studied the language of prose and made the first study of syntax, the first vocabulary of the written language, and a classification of the so-called figures of speech. About the same time Didymus, known as Chalcenterus ("Brazen-Gutted"), incorporated into huge variorum editions much of the precious material contained in the many commentaries on literature compiled during the Hellenistic Age. This vastly productive scholar was lacking in critical judgment, but it is on his work that the later less extensive commentaries that in part survive depended. Under Tiberius, Theon studied the Hellenistic poets, as well as Pindar.

The Attic Revival

The 1st century AD saw the beginning of the "Attic Revival," the movement to imitate the language and style of the classical Athenian writers, which lasted far into the Byzantine period with disastrous effects that have not even yet died away. This resulted in the production of many lexica and manuals meant to help people to write correct Attic, such as the works of Phrynichus, Moeris, and Pollux, all probably dating from the 2nd century AD. At that time much learned work was still being done, but it was becoming increasingly mechanical and repetitive. More and more of the chief writers survived only in selections; texts were being produced, often with commentaries, but these derived mainly from the stores of learning accumulated in the past. However, under Hadrian, Apollonius Dyscolus produced a treatment of syntax that acquired great authority, and his son Herodianus produced the standard treatise on accentuation; they were the last known producers of important original work on grammar.

The rise of Christianity

Christianity proved less hostile to pagan culture than might have been expected. From the 2nd century on, Church Fathers such as Justin, Clement of Alexandria, and Origen used an impressive knowledge of pagan literature to debate with pagan philosophers on equal terms. Prominent on the pagan side was the Neoplatonist Porphyry (*c.* 234–*c.* 305). Besides his published attacks on Christianity, he wrote commentaries on Plato, Aristotle, Theophrastus, and Plotinus. Even after the triumph of Christianity in 313 under Constantine the Great, pagan and Christian scholars often attended one another's lectures. The pagan Libanius of Antioch, the most celebrated rhetor of the time and author of the surviving hypotheses of the orations of Demosthenes, taught Theodore of Mopsuestia, St. John Chrysostom, and probably also St. Basil and Gregory of Nazianzus. Basil (*c.* 329–379) wrote a treatise on the value of pagan literature in which he recommends at least a passing acquaintance with the pagan classics, but he and the other leading Christian authors of his time possessed a good deal more than this. Theodore (*c.* 350–428/429), bishop of Mopsuestia and leader of the school of Antioch, applied what could be called pagan methods of criticism to the Bible by using his knowledge of history and language to illuminate passages of Scripture. Members of the Christian school of Gaza in the 5th and 6th centuries even wrote dialogues modeled on those of Plato. The school's leading member, Procopius, invented the *catena* ("chain"), a commentary on a book of the Bible consisting of a compilation of excerpts from earlier commentaries—something obviously suggested by the variorum editions of classical authors. Notes based on the learned commentaries of the Hellenistic Age now came to be written into the margins of manuscripts; to these scholia is owed most of what is known of ancient scholarship.

The Neoplatonists of the 5th and 6th centuries produced commentaries on Plato, Aristotle, and other philosophers, thus preserving many priceless fragments of earlier philosophical texts now lost. Grammatical work also continued: Proclus wrote a commentary on Hesiod's *Works and Days;* Hesychius of Alexandria compiled a Greek lexicon that preserved vocabulary from the Homeric age up to his own time; and Orus contributed to the work on Greek orthography. Education even received some government support; the 4th-century rhetor Themistius described a plan for the creation of a government scriptorium to ensure the survival of important writers, and some 50 years later, in 425, Emperor Theodosius II is said to have set up a university at Constantinople.

The age of Justinian I (527–565) produced the antiquarian works of Johannes Lydus and the geographical gazetteer of Stephanus of Byzantium. The historians of that era, Procopius and Agathias, wrote in the classical tradition of historiography, publishing chronicles of warfare that weighed the influences on historical events of fate and divine retribution. But in 529 Justinian issued an edict closing the schools of pagan philosophy; some philosophical activity continued after that, but the edict

marked an era of Christian intolerance of pagan scholarship. During the 7th century the Arab conquests cut off Syria, Palestine, and Egypt from Greek civilization. The Arab threat forced the Byzantine Empire to submit to the rule of vigorous but not well-educated emperors, some of whom were religious fundamentalists opposed to the use of images, or icons, which was a central feature of worship in the Eastern Church. The resulting Iconoclastic Controversy was a major factor in the creation of a dark age of Byzantine culture that lasted from about the middle of the 7th until the beginning of the 9th century.

The first Byzantine renaissance. The dark age was not completely dark. It saw, for example, the extensive but exceedingly uninspired work of the grammarians Georgius Choeroboscus, active during the second half of the 8th century, and Theognostus, early in the 9th century, as well as the letters of the deacon Ignatius with their surprising wealth of literary allusions. Also, certain developments that occurred at this time were important for the future. In about 800, paper was acquired from the Arabs, who are said to have learned how to make it from Chinese prisoners taken in a battle at Samarkand. It came into general use only very gradually; the Byzantines continued to import it from the Arabs instead of making their own, but since it was less expensive than papyrus, its effect was bound to be important. The Italians acquired it from the Byzantines, and by the 13th century they had developed a flourishing paper industry. From about the same time must date the invention of a new cursive script, the Byzantine minuscule, which was in its early forms the most elegant that the Greeks ever invented. The earliest surviving specimen, the Uspenskij Gospel, dates from 835, but this displays such accomplished writing that the new script probably originated some 50 years earlier. The invention greatly facilitated the rapid production of books. The Stoudion monastery in Constantinople, which flourished under its great abbot St. Theodore (759–826), was once thought to have introduced the new script—and indeed the monastery had a flourishing scriptorium—but this conjecture is by no means certain. During the 9th and 10th centuries the works of many classical authors were transferred from manuscripts in the old uncial writing to the new minuscule, and the surviving books of this period show that script in its most perfect form. Later the elegance of minuscule was spoiled by the admixture of uncial letters and the increasing use of ligatures.

Byzantine minuscule script

The first important scholar of the first Byzantine renaissance was Leo the Philosopher (*c.* 790–*c.* 869), a notable teacher in Constantinople who numbered among his pupils St. Cyril, one of the apostles of the Slavs; Leo had considerable knowledge of Greek culture, particularly of science and mathematics. But the dominant figure in the revival of the 9th century was the patriarch Photius (*c.* 820–891?), who not only compiled a notable Greek lexicon but also produced the *Myriobiblon,* or *Bibliotheca,* a vast collection of summaries and evaluations of various ancient books, mainly historical. Photius also compiled a learned miscellany called the *Amphilochia* and an interesting collection of letters. Arethas (born *c.* 850), archbishop of Caesarea Cappadociae, owned a remarkable private library, from which eight priceless books, commissioned from the finest calligraphers of the time, survive; Euclid, Plato, Aristotle, Lucian, and Aristides are among them. Other valuable classical manuscripts still extant formed part of his collection.

During the 10th century education was encouraged by the learned emperor Constantine VII Porphyrogenitus (905–959), who apart from producing his own series of historical works preserved several histories by others and planned a vast 53-section encyclopaedia of human activities that was probably never completed. The 10th century also saw the production of a large encyclopaedia cum dictionary, formerly thought to have been the work of one Suidas, but now known to have been called the *Suda,* from a Byzantine Greek word for fortress. Platonism was actively studied by the chief intellectual figure of the 11th century, Michael Psellus (1018–*c.* 1078). His numerous writings show a wide acquaintance with classical culture, though also a very imperfect sympathy with some of its elements. His pupil, John Italus, was anathematized by the ecclesiastical authorities for allowing Platonism to contaminate his Christianity. But Platonic studies continued, and Isaac Sebastocrator, a brother or son of the emperor Alexius I Comnenus, wrote three essays based on Proclus. Early in the 12th century Alexius' daughter, Anna Comnena, was the centre of a circle of Aristotelian scholars, including Michael of Ephesus and Eustratius, who together produced a commentary on the *Ethics.* Gregory of Corinth, active during the same period, wrote works on syntax and style and also one of the few ancient treatments of the Greek dialects that have come down to the present. John Tzetzes wrote some 60 books on Greek literature that are learned but uncritical, and Eustathius of Thessalonica wrote vast commentaries on the *Iliad* and *Odyssey* that incorporate much earlier learning.

This epoch of Byzantine learning was rudely put to an end when the knights of the Fourth Crusade, under Venetian leadership, sacked Constantinople in 1204. It may well be argued that this event was an even greater disaster for learning than the Turkish capture of the city in 1453, for which the crusaders paved the way. The sack of the city destroyed a quantity of Greek literature that is difficult to estimate; certainly included among the lost works were the *Aitia* and *Hekale* of Callimachus, which were known to Michael Choniates, archbishop of Athens at the time of the Crusade.

Losses during the Fourth Crusade

The second Byzantine renaissance. Between 1204, when the imperial capital was moved to Nicaea, and 1261, when Constantinople was recovered by the Palaeologus dynasty, classical studies continued under the difficult conditions outlined in the autobiography of Nicephorus Blemmydes, the leading intellectual of the time. The emperor Theodore II Lascaris (reigned 1254–58) did much to assist cultural life during this time. The period sometimes called the Palaeologan Renaissance saw a revival of classical studies that, under the circumstances, must be called remarkable. Maximus Planudes (1260–*c.* 1310) made many compilations, including a new anthology of epigrams, and even translated into Greek such Latin texts as parts of Ovid, Augustine, and Boethius. He also had some knowledge of Hellenistic poetry and even Arab astronomy and mathematics. From about 1300 the texts of the Greek dramatists were studied critically by Manuel Moschopoulos, Thomas Magister, and finally Demetrius Triclinius. Triclinius had read the metrical handbook of the 2nd-century scholar Hephaestion and understood the simpler metres, and he was also aware of the principle of strophic responsion. He was therefore able to make a number of emendations worthy of serious notice. Theodore Metochites (*c.* 1260–1332), one of the leading intellectuals and public men of his time, commented on Aristotle and wrote a miscellany that contains interesting reflections upon classical authors, especially orators and historians.

Greek in the West. During the 3rd and 4th centuries the knowledge of Greek in the West died out with shocking suddenness; Augustine had only a rudimentary knowledge of the Greek language, and translators such as Jerome (*c.* 347–419/420) and Rufinus (*c.* 345–410/411) were scarce indeed. The few Greek studies were undertaken for the sake of theology or philosophy, and translation of secular authors was rare; Calcidius' (Chalcidius') 4th-century version of the *Timaeus* was for eight centuries the only Latin translation of a Platonic dialogue, Boethius' plan for a series of translations of Plato and Aristotle being interrupted by his execution. Sicily remained Byzantine until the Arab conquest of the 9th century, and Calabria, Lucania, and Apulia (Puglia) until the Norman conquest of the 11th century. The Normans and later the Hohenstaufen rulers favoured Greek studies. In the 12th century Greek, too, benefited from the intellectual revival; Henricus Aristippus, archdeacon of Catania, translated Plato's *Meno* and *Phaedo,* and the admiral Eugenius collaborated in a Latin version of the *Almagest,* an encyclopaedia compiled by the astronomer Ptolemy of Alexandria in the 2nd century AD. Also during the 12th century two Italian scholars, James of Venice and Burgundio of Pisa, traveled to Constantinople in search of theological and philosophical learning; Burgundio brought back literary as well as theo-

Aristotelian revival

logical manuscripts, though he was probably incapable of reading them. The Aristotelian revival of the 13th century led to the production of many translations of Aristotle by William of Moerbeke in Rome, and in England Aristotle was read in the original by Robert Grosseteste and Roger Bacon. During the 14th century contact between Rome and Constantinople was continued; Petrarch (see below *Latin scholarship*) acquired a Byzantine manuscript of Homer, though he never made the effort to enable himself to read it, and later in the century another such manuscript was in the hands of the humanists of Padua. In about 1397 the Byzantine scholar Manuel Chrysoloras went to Italy to teach Greek in Florence. At the Council of Ferrara-Florence in 1438–45 the union of the churches was agreed upon, but it was later repudiated. George Gemistus Plethon (*c.* 1355–1450/52), the famous Neoplatonist of Mistra, was present at that council; with him was his pupil John Bessarion of Trebizond (1403–72), who continued to support church union as an individual, so that when the repudiation took place he converted to the Western church. He stayed behind in Italy, became a cardinal, and made an important gift of books to Venice. Early in the 15th century Italians such as Francesco Filelfo and Giovanni Aurispa were bringing back Greek manuscripts from Constantinople in large quantities, so that well before the capture of Constantinople by the Turks in 1453 many Greek books had found their way to the West.

Latin scholarship. *Republic and early empire.* From the beginning, Roman scholarship imitated Greek: Hellenistic techniques were applied to the treatment of Latin texts, and Latin grammar adopted Greek categories and terminology. Learned Greeks such as Tyrannion, Alexander Polyhistor, and Parthenius were brought to Rome as prisoners in the Mithradatic Wars. Even before that, as early as about 100 BC, the Roman knight Lucius Aelius Stilo Praeconinus had been teaching and writing about Latin grammar. Marcus Terentius Varro (116–27 BC) by his vast learning and prodigious output influenced almost every branch of scholarship; of his 25 books about the Latin language, books v to x survive in nearly complete form. In scholarship as in other matters the early imperial period was one of great achievement. It was the age of commentators such as Gaius Julius Hyginus, who was in charge of the Palatine Library in Rome founded by Augustus; of editors such as Marcus Valerius Probus (*c.* AD 20–105), who made critical editions of Plautus, Terence, Lucretius, Virgil, and Horace; of grammarians such as Verrius Flaccus, the author of a vast work on the meaning of words; of the elder Pliny (AD 23/24–79), whose encyclopaedic *Historia naturalis* (*Natural History*) was a major sourcebook during the Middle Ages; of Gaius Suetonius Tranquillus (*c.* AD 69–after 122), who wrote the lives of poets and grammarians as well as of emperors; and of Aulus Gellius, whose miscellany called *Noctes Atticae* preserved much ancient learning.

Later empire. The barbarian invasions of the 3rd century marked the beginning of a testing time for Latin as well as for Greek scholarship, and the scholars of the 4th and 5th centuries—such as Aelius Donatus, the grammarian and teacher of rhetoric; Servius, the learned commentator on Virgil; Priscian, the Greek author of the most famous Latin grammar of antiquity; and Macrobius, who during the first half of the 5th century wrote the learned miscellany called *Saturnalia*—were epitomizers and compilers living on inherited capital. In the Western Empire the knowledge of Greek was practically extinct, and the earlier literature of Rome itself was threatened with extinction. The classics were still the staple of such education as there was, but the dominance of rhetoric favoured only certain authors; classical poets such as Virgil, Horace, Ovid, and Terence were protected by critics and commentators, but among earlier authors Ennius and Lucilius disappeared and Plautus narrowly escaped. The book, in the form of the vellum or parchment codex, was superseding the papyrus roll, and authors who were not recopied were doomed to oblivion.

The dominance of rhetoric

The early Middle Ages. The period during which the Merovingian dynasty founded by Clovis (*c.* 466–511) was in power was a dark age for learning, but there was no complete breach with the past. Under the influence of the church the barbarian invaders wished to base their civilization on the Latin model, and since it was the language of the church, Latin continued to be the language of literature. Although interest in antiquity for its own sake had little part in the late imperial and early medieval ideal, under the protection of the church learning survived in the medieval schools, and classical texts provided a grounding in grammar, a training in logical thought, and a philosophical premise for theology. Flavius Cassiodorus, a retired statesman who founded a monastery at Vivarium, in southern Italy, sometime after AD 540, encouraged his monks to copy pagan as well as Christian authors, a practice that spread later to other monasteries, particularly those of the Benedictine order. About 563 the Irish missionary Columba (*c.* 521–597) founded a church and monastery on the island of Iona in the Inner Hebrides of Scotland, and soon afterward Irish missionaries converted the whole of Scotland and established monasteries in the north of England. Later Irish missions led by Columban (*c.* 543–615) founded Luxovium (Luxeuil) in the Vosges Mountains of Gaul (590), Bobbio on the Trebbia (*c.* 612–614), and St. Gall in Switzerland; Corbie near Amiens was founded from Luxovium a century later, and these monasteries played a leading role in the preservation of ancient literature. In England, Aldhelm (*c.* 639–709) and Bede (672/673–735) were men of considerable learning. At this time learning was also alive in Visigothic Spain, as is shown by the vast encyclopaedia of Isidore of Seville (*c.* 560–636), which was of great importance for the remainder of the Middle Ages. During the late 7th and 8th centuries the successors of Columban converted first the Frisians and then much of Germany; they founded the important monasteries of Fulda (744), Lorsch, in Hesse (764), and Hersfeld (*c.* 770), while Reichenau, on Lake Constance (724), and nearby Murbach (727) were founded by a refugee from Visigothic Spain.

Christianization of the British Isles

The Carolingian Renaissance. Pepin III the Short (reigned 751–768) began ecclesiastical reforms that Charlemagne continued, and these led to revived interest in classical literature. Charlemagne appointed as head of the cathedral school at Aachen the distinguished scholar and poet Alcuin of York, who had a powerful influence on education in the empire. Many ancient texts were now copied into the new Carolingian minuscule, and the palace library allowed its books to be copied for other libraries, so that learning was rapidly diffused. Latin poetry of some merit was composed at and about the imperial court, and Einhard's life of Charlemagne (probably written *c.* 830–833) is modeled on the biographies of Suetonius. Learned work was resumed, and the historian Paul the Deacon (Paulus Diaconus) abridged the abridgement of the lexicon of Verrius Flaccus that had been made by Festus during the 2nd century AD. The nearest approach in the Middle Ages to a humanistic scholar was Servatus Lupus, abbot of Ferrières (*c.* 805–862), who collected, copied, and excerpted ancient manuscripts on a large scale. Despite the splitting up of the Carolingian Empire in 843 and the troubles resulting from the barbarian attacks on Europe of the 9th and 10th centuries, the educational apparatus created by the so-called Carolingian Renaissance provided enough momentum to keep the classical tradition going until a new impulse arrived to carry it on to fresh developments.

Carolingian minuscule

The later Middle Ages. A renewed period of intellectual activity in the ancient Benedictine foundation of Monte Cassino heralded the renaissance of the 12th century. Dante Alighieri (1265–1321) was familiar not only with Virgil but also with Lucan, Statius, and Ovid, and *The Divine Comedy*'s picture of the cosmos is deeply indebted to Aristotle's *On the Heavens*. William of Malmesbury (died *c.* 1143) and John of Salisbury (1115/20–1180) were considerable Latin scholars. During the 13th century a group of scholars in Padua around Lovato Lovati (1241–1309) and Albertino Mussato (1261–1329) were active humanists. Lovato read Lucretius and Catullus, studied Seneca's tragedies in the famous *Codex Etruscus*, and found and read some of the lost books of Livy. Both men wrote Latin poetry, Mussato composing a Senecan tragedy, the *Ecerinis*, designed to open the eyes of the Paduans to the

12th-century renaissance

danger presented by Cangrande della Scala, the tyrant of Verona, by describing the tyrannical conduct of their own former despot, Ezzelino III.

THE REVIVAL OF LEARNING

Renaissance humanism. The humanist movement was consolidated by the generation of Petrarch (Francesco Petrarca; 1304–74). Petrarch actively looked for manuscripts, building up what was for his day a remarkable library, and taught himself to write an elegant classicizing Latin very different from what had been customary during the Middle Ages. Like Politian later, he was a great poet in Italian; but he valued far more than his vernacular poetry his Latin epic *Africa,* a skillful imitation of the Roman poets. Like almost everyone before Politian, Petrarch knew little or no Greek (on the manuscript of Homer that he possessed, see above, *Greek in the West*). Giovanni Boccaccio (1313–75) also looked actively for ancient manuscripts and actively forwarded the aims of humanism.

The age of Petrarch

The revival of classical learning that Petrarch and Boccaccio promoted was only one aspect of the complex phenomenon of the Renaissance. In origin the movement was utilitarian, seeking to exploit classical antiquity in the service of modern man; the early Italian humanists were not scholars so much as *litterati* and educators, and it is a mistake to think that they were pagans. The earlier idea that the invention of printing was an effective agent in the revival is also erroneous; for by 1470, when the first editions of the Latin classics were quickly coming off the presses, the Renaissance was already well past its early stages. Thus, although Greek teachers and Greek manuscripts had long before begun to enter Italy, the advanced study of Greek, apart from the activities of an isolated genius like Politian, made little headway before the 16th century. The early humanists saw that the manuscripts they discovered contained many corruptions and enjoyed trying to emend them, but many of their conjectures were frivolous, and they often omitted to mark them as conjectures, a practice that irritated later scholars.

Petrarch's successor as the leader of the humanist movement was Coluccio Salutati (1331–1406), chancellor of Florence, who acquired many manuscripts and built up a splendid library; it was he who invited Chrysoloras to Florence. A later chancellor, Leonardo Bruni (*c.* 1370–1444), translated into Latin Plutarch, Xenophon, six dialogues of Plato, and Aristotle's *Ethics* and *Politics.* Poggio Bracciolini (1380–1459) was the most active and the most successful hunter of manuscripts, traveling to France, Germany, and even England in pursuit of them. The same period saw the beginning of the study of the ancient monuments of Italy and the collection of coins and inscriptions as well as works of art by scholars such as Flavio Biondo (Flavius Blondus; 1392–1463) and later Pomponius Laetus (1428–97). Cyriacus of Ancona (1391–1452) broke new ground by traveling to the countries of the Turkish Empire, where he drew monuments and copied inscriptions, thus providing the only record of many objects that were later lost.

Beginnings of modern scholarship. What may be called professional standards of scholarship are seen first in the work of Lorenzo Valla (1407–57) and Politian (Angelo Poliziano; 1454–94). Valla in his *Elegantiae* demonstrated the technique of pure and elegant classical Latin, free of medieval awkwardness; when Pope Nicholas V ordered the chief Greek prose writers to be translated into Latin, Valla was responsible for Thucydides. He also translated part of the *Iliad* into Latin prose. In his philosophical works, which include treatises on pleasure and on free will, he was the first modern to throw light on Epicurus. Gifted with the historical sense of the true critic, Valla perceived the spuriousness of several famous documents: a treatise forged in the name of Dionysius the Areopagite, the New Testament convert of St. Paul, by a later writer now called Pseudo-Dionysius; a collection of letters supposedly exchanged by St. Paul and the 1st-century-AD Roman philosopher Seneca; and the so-called Donation of Constantine, by which the emperor Constantine the Great was alleged to have granted to the papacy spiritual and temporal dominion over Rome and the West.

Politian, like Petrarch a great poet in the vernacular, began studying Greek at the age of 10 and attained a better knowledge of it than any modern to that date; in his collection of notes called *Miscellanea,* the second volume of which was unfortunately lost and was published only in 1972, he threw light on a variety of ancient writers, including even Greek poets of the Hellenistic Age.

Politian

By 1500 most of the chief Latin authors were in print. In that year Aldus Manutius (1449–1515) founded in Venice his "Neacademia" (or Aldine Academy), dedicated to, among other things, the issuing of large and relatively cheap editions of ancient authors. Working in conjunction with the learned Cretan Marcus Musurus (1470–1517), he brought out in 21 years 27 editiones principes (first editions) of Greek authors, including five in the year 1502 alone. During the century that followed, the book evolved from what was essentially an expensive facsimile of a medieval manuscript into a working tool for scholars. Other printers, such as the Giunta family in Florence, followed Aldus' example, and Zacharias Callierges in Rome brought out the first printed texts of Pindar, Callimachus, and the Homeric scholia. Aldus' son Paulus Manutius (1512–74) carried on his father's business and did much for the texts of Cicero. Petrus Victorius (1499–1585) was the leading Italian scholar of his time, editing Aeschylus and Euripides and writing commentaries on Aristotle's *Rhetoric, Poetics, Politics,* and *Nichomachean Ethics,* as well as editing other Greek texts and doing important work on Cicero; he concentrated on producing careful editions of the best manuscripts available, in a reaction against the excessive emendation of earlier scholars. Francesco Robortello (1516–67) also did important work on Aeschylus and Aristotle's *Poetics.* Fulvius Ursinus (1529–1600) built up the Farnese library in Rome, edited the Greek lyric poets, and made important contributions to numismatics and iconography. Carolus Sigonius (1523–84) and Pirro Ligorio (*c.* 1510–83) were active in the field of history and antiquities, Ligorio producing much genuine material besides his notorious forgeries. But after the 16th century, the atmosphere of the Counter-Reformation was not favourable to disinterested inquiry, and Italian scholarship declined. The Jesuits in their educational activities made use of the forms of humanism while abolishing its content.

Advancement of the book

The Renaissance outside Italy. In Spain the Renaissance had made a promising beginning; Antonio of Nebrija (1444–1522) anticipated Erasmus in showing that the Greek language had been pronounced by the ancients differently from the modern Greeks, and later Antonio Agostino, archbishop of Tarragona (1517–86), did important work on ancient law and numismatics. But the Spanish Renaissance was frozen by the Counter-Reformation.

During the late 15th and early 16th centuries the new learning began to establish itself north of the Alps. William Grocyn, who had studied in Italy, was probably the first man to teach Greek in an English university; he was friendly with John Colet and Thomas More, both of England, and later with the Dutch humanist Desiderius Erasmus. Thomas Linacre, later an eminent physician, studied Greek in Italy under Politian; on his return to England he gave lectures at which More was present.

The Northern Renaissance

Erasmus (*c.* 1466–1536), the first editor of the New Testament, was more concerned with biblical and patristic studies than with the Greek and Latin classics for their own sake. Yet his *philosophia Christi,* an attempt to mediate between ancient wisdom and Christian faith, was closely linked with classical scholarship, and he found time to produce numerous editions and translations of Greek and Latin authors, besides making such contributions to scholarship as his famous collection of proverbs, the *Adagia.* The *Utopia* of his English friend Thomas More was profoundly influenced by Platonism. Erasmus' pupil Beatus Rhenanus was one of a group of German scholars who brought out important editions of Latin texts. Philipp Melanchthon (1497–1560) actively promoted scholarship in Germany; his associate Joachim Camerarius (1500–74) did much for Plautus, as did Hieronymus Wolf (1516–80) for the Attic orators.

Erasmus formed a close connection with the great printer

of Basel, Johannes Froben. Amerbach, Cratander, and Hervagius were other notable printers of that city, active in the production of critical editions of ancient texts.

Obliged as they were to concede primacy in Latin studies to the Italians, the French during the 16th century took the lead in Greek, although Denis Lambin (Lambinus; 1516–72) did valuable work on Cicero, Lucretius, and Horace. Guillaume Budé (Budaeus; 1467–1540) laid the foundations in Greek studies, and Jean Dorat (Auratus; 1508–88) and Adrien de Tournebu (Turnebus; 1512–65), pioneers in the study of Greek poetry, inspired the contemporary poets Ronsard and du Bellay, the leaders of the Pléiade group, with admiration for Greek literature. The great printer Robert Estienne (Stephanus; 1503–59) produced the first critical edition of the Greek New Testament (1551), reprinting Erasmus' text but adding variants from 15 manuscripts. Estienne's son, Henri, published many editiones principes of Greek authors and a Greek *Thesaurus* (1572) not superseded until the early 19th century.

Scaliger and Casaubon

Two French scholars—Joseph Justus Scaliger (1540–1609) and Isaac Casaubon (1559–1614)—deserve particular mention. Like Erasmus, Scaliger saw that classical learning should be a unity. His diversity was that of the explorer, not the dilettante; each edition opened up a new path: that of Festus (1575) to Old Latin, that of Manilius' *Astronomica* (1579) to ancient astronomy, for example. He assisted Janus Gruterus (1560–1627) by compiling the indexes to his famous *Inscriptiones antiquae totius orbis Romani* and encouraged the collection of the fragments of classical literature. But his greatest achievement was to bring order into the chaos of ancient chronology in his *De emendatione temporum* (1583) and *Thesaurus temporum* (1606).

Casaubon, too, perceived that antiquity must be studied as a whole and also (and this too Erasmus understood) that the study must begin from Greek. Through his series of detailed commentaries on difficult and prolific authors (Strabo, Athenaeus, Polybius), he was instrumental in turning scholarship—hitherto an art—into a science.

Henri Estienne, Scaliger, and Casaubon were all Huguenots, and all died in exile—Estienne in Lyon, Scaliger in Leiden, and Casaubon in London. Another eminent Huguenot scholar of the time, Marcus Antonius Muretus (Marc-Antoine de Muret; 1526–85), the most elegant writer of Ciceronian Latin since Cicero, who defended the practice of emendation against the cautious Victorius, left France when accused of homosexuality, became a Catholic, and enjoyed great success in Rome.

Scholarship in the 17th century. After the conversion of Henri IV to Roman Catholicism French scholarship declined, as Italian scholarship declined during the age of the Counter-Reformation. But the action of the Jesuits in challenging the authenticity on which the privileges of the Benedictines depended caused the latter to turn to the study of paleography in order to defend themselves, thus occasioning the chief contribution of France to classical studies during the 17th century. Jean Mabillon (1632–1707) established Latin paleography as a modern science, and another inmate of the monastery of St. Germain-des-Prés, Bernard de Montfaucon (1655–1741), did the same for Greek paleography. This kind of work was continued by the great antiquarians of the following century, notably L.A. Muratori (1672–1750) and Scipione Maffei (1675–1755).

Development of paleography

As scholarship declined in France (where the series of Delphin Classics supervised by Pierre-Daniel Huet from 1670 to 1680 marks the summit of strictly classical achievement), so it rose and flourished in the Netherlands. Christophe Plantin had founded his great press in Antwerp in 1550 and the Elzevirs theirs in Leiden in 1580 and later in Amsterdam. Scaliger ended his days in the newly founded State University of Leiden. Justus Lipsius (1547–1606) produced important editions of Tacitus and Seneca, at the same time promoting a new Christian Stoicism. The great jurist Hugo Grotius (1583–1645), in many ways the true successor of Erasmus, did brilliant work in classical studies as well as in many other fields. Nicolas Heinsius (1620–81) produced editions, based on extensive study of manuscripts, that earned him the title "saviour of the Latin poets." His counterpart in prose, John Gronovius (1611–71), produced editions of Livy, Seneca, Pliny, and others. The letters of Heinsius and Gronovius testify to as ample a conception of classical studies as that of Scaliger and Casaubon.

Impact of the Thirty Years' War

The Thirty Years' War (1618–48) had a disastrous effect on scholarship. Among the classicists in Holland and Germany, stagnation set in; unwieldy, uncritical variorum editions became the fashion, and the collection of "antiquities," divorced from linguistic study and from critical scholarship, degenerated into the mere piling up of information.

The 18th century: the age of Bentley. Since the late 16th century little had been heard of English scholarship; once the study of Greek had been established by Linacre, Grocyn, Sir John Cheke, and their contemporaries, the English preoccupation with education had set in. John Selden is the most notable of few exceptions, and he was a jurist and antiquary, not an academic, though his *De Diis Syris* (1617) laid the foundations of Eastern scholarship. A new era began with the *Epistola ad Joannum Millium* (1691) of Richard Bentley (1662–1742). This collection of brilliant miscellaneous observations, prompted by the editio princeps of the 6th-century Byzantine chronicle of John Malalas, displayed already the comprehensive learning and rare power of divination that were to enable Bentley to lay the foundations of the critical scholarship of the coming age. Although his achievements in textual criticism were singularly brilliant, Bentley must not be thought of as a mere editor of texts but as the creator of a critical method that was to be applied with powerful effect in every department of antiquity. This is in evidence above all in his *Dissertation upon the Epistles of Phalaris* (expanded edition, 1699), the first important work of classical scholarship written in a modern language. His editions of Horace (1711), Terence (1726), and Manilius (1739) were all of masterly quality. He did remarkable work in collecting fragments of Menander and Callimachus, and although he never completed his proposed editions of Homer and the New Testament, the preparatory work he did toward them had a revolutionary effect in both fields of study.

After Bentley's death the only part of his inheritance taken up by his countrymen was his work in textual criticism. The work of his English contemporaries in this field, who include such important scholars as Jeremiah Markland (1693–1776), Thomas Tyrwhitt (1730–86), Benjamin Heath (1704–66), and Samuel Musgrave (1732–80), was carried further by the next generation. Richard Porson (1759–1808), Peter Elmsley (1773–1825), and P.P. Dobree (1782–1825) all concentrated upon Attic drama, Porson showing a particularly fine feeling for Greek.

In 1786 Sir William Jones (1746–94) began the study of Sanskrit that was to lead to the establishment of the new discipline of comparative philology. Edward Gibbon (1737–94), essentially self-educated despite his early residence at Magdalen College, Oxford, made with *The History of the Decline and Fall of the Roman Empire* (1776–88) the greatest single contribution to the study of ancient history in the whole 18th century. The *Essay on the Original Genius of Homer* by Robert Wood (*c.* 1717–71), printed privately in 1767 and published posthumously in 1775, not only marked a new stage in Homeric studies but also assisted the movement toward exploration of ancient sites in Greece. Exploration was powerfully promoted by the publications in London of the Society of Dilettanti, especially the drawings in *The Antiquities of Athens* (four volumes, 1762–1808), by James Stuart and Nicholas Revett.

Edward Gibbon

Meanwhile in the Netherlands, where Bentley's greatness had at once been recognized, a distinguished series of scholars—Tiberius Hemsterhuys (1685–1766), L.K. Valckenaer (1715–85), the German emigrant David Ruhnken (1723–98), and, later, Daniel Wyttenbach (1746–1820)—continued to do valuable work on Greek texts, including the difficult but rewarding remains of ancient lexicography. Bibliographical works and dictionaries were now improved; Johann Albert Fabricius (1668–1736) put the bibliography of Greek and then Latin literature on a new footing, and Egidio Forcellini in Padua superseded the

Latin thesaurus of Robert Estienne. The study of ancient coins was greatly advanced by the work of the Swiss-born scholar Ezechiel Spanheim (1629–1710) and the Austrian scholar J.H. Eckhel (1737–98).

In archaeology the 18th century saw the beginning of the excavation of Herculaneum and Pompeii and of exploration of the remains of the Etruscan civilization. Historical source criticism (*Quellenkritik*) began in the work of the German historian Barthold Niebuhr (1776–1831). But technical progress was not enough. A new spirit was needed to arouse classical studies to take their place in the modern world, and it came from Germany.

MODERN CLASSICAL SCHOLARSHIP

The new German humanism. The "new humanism" that transformed German intellectual life in the late 18th century was a complex phenomenon, acting through scholarship, education, philosophy, and literature. Educationally the University of Göttingen played a leading part: there J.M. Gesner (1691–1761) and C.G. Heyne (1729–1812) introduced a new approach—an attempt to enter into the spirit of the past as displayed in its artistic monuments as well as in its literature. J.J. Winckelmann (1717–68) was the first to mark out the successive periods into which the history of Greek art falls. He was also the first to isolate and describe the essentially Hellenic element in Greek art and to relate the development of art in antiquity to other aspects of culture. He demonstrated that a large number of vases then known as Etruscan because they had been found in Etruscan cemeteries were in fact Greek, although the original error was to be perpetuated by Josiah Wedgwood, who named his pottery works "Etruria" in 1769. Winckelmann's influence ranged over the literary as well as the academic world, powerfully affecting such figures as Lessing, Herder, and Goethe, who named the memorial essay that he published in 1805 *Winckelmann und sein Jahrhundert* ("Winckelmann and His Century"). Goethe (1749–1832) made a systematic effort to know and understand Greek art and literature, particularly Greek sculpture and the poetry of Homer, and to utilize them for his own purposes; it may be doubted whether anyone since ancient times had understood the Greeks so well. He was a friend of Wilhelm von Humboldt (1767–1835), the great statesman and pioneer of the study of language who founded Frederick William University (later the University of Berlin), which rapidly became the leading university of Europe.

The achievement of Goethe

Goethe was also in touch with F.A. Wolf (1759–1824), a Göttingen pupil of Heyne. Wolf defined the "science of antiquity" (*Altertumswissenschaft*) and mapped out its constituent provinces and principles. Influenced by Herder, with his special interest in the early literatures of various peoples and their special characteristics, Wolf in his *Prolegomena ad Homerum* (1795) raised such questions about Homer as to give rise to a debate that has continued ever since. Goethe was at first carried away by Wolf's theory that the *Iliad* and *Odyssey* were composed orally by a number of authors and that the artistic unity of the poems was a later imposition, but he eventually returned to his belief in an individual Homer, as scholars have done increasingly in the 20th century.

Another great classical scholar in close touch with Goethe was Gottfried Hermann (1772–1848), who continued the tradition of 18th-century rationalism, applying to the study of ancient poetry a critical method based on a strict Kantian logic. Hermann did much to advance the study of Homer, Pindar, late epic poetry, and Greek metre. By his editions of Sophocles, Aeschylus, and part of Euripides he effected a great and permanent improvement in the texts of these poets.

Hermann had many distinguished pupils, including C.A. Lobeck (1781–1860), a grammarian of great learning and acuteness, who in his famous book *Aglaophamus* (1829) refuted the seductive but dubious theory of the Heidelberg professor G.F. Creuzer that the mythology of Homer and Hesiod contained symbolic elements of an ancient Oriental revelation from which it was ultimately derived. August Meineke (1790–1870) did important work on Hellenistic poetry and produced an excellent edition of the fragments of Greek comedy. August Immanuel Bekker (1785–1871), a pupil of Wolf, took advantage of the accumulation in Paris of many previously inaccessible manuscripts from various countries following the Napoleonic conquests to make a valuable contribution to the texts of many prose authors. Wilhelm Dindorf (1802–83) edited many texts, including the scholia on Rome and on Demosthenes. With his brother Ludwig and K.B. Hase, he revised the great Greek *Thesaurus* of Henri Estienne. Hermann's son-in-law Moritz Haupt (1808–74) did important work on Latin poetry. H.L. Ahrens (1809–81) wrote on the Greek dialects and on the bucolic poets. August Nauck (1822–92), who taught in St. Petersburg, made a notable contribution to the establishment of the texts of Greek tragedy.

The school of Hermann with its strong emphasis on linguistic study came occasionally into conflict with the representatives of a newer trend in the approach to antiquity. In Berlin August Boeckh (1785–1867) did important work on Greek poetry, particularly Pindar, but also established on a firm footing the study of Greek private and public economy and the systematic collection of Greek inscriptions. K.O. Müller (1797–1840), the author of an important history of Greek literature, which first appeared in English, was a pioneer of the study of Greek, Roman, and Etruscan origins and of the mythology of the different parts of Greece, which he believed could shed much light on early history. He was a strong upholder of the importance of art and archaeology in the study of antiquity, as was F.G. Welcker (1784–1868), who applied deep knowledge of Greek art and religion to the interpretation of literature and did much to shape the wider conception of the study of antiquity that was now coming to maturity.

Comparative linguistics

The comparative study of Indo-European languages that was initiated by Franz Bopp (1791–1867), one of the famous scholars who gave the University of Berlin its enviable reputation, profoundly influenced the study of the ancient as well as other languages. One field in which this was seen was the study of early Latin, which was now placed on a new basis by Friedrich Ritschl (1806–76), who applied knowledge gained from the study of inscriptions to the elucidation of early Latin texts. There followed much important work on early Latin, such as that of Johannes Vahlen (1830–1911) on Ennius and that of Otto Ribbeck (1827–98) on Roman tragedy.

The rise of textual criticism. Knowledge of ancient literature must always rest on the standards of editing and criticism of Greek and Latin texts that have come down in a corrupt and sometimes mutilated state. Early in the 19th century great advances were made in this field of classical studies. Angelo Cardinal Mai (1782–1854) published hitherto unknown Greek and Latin texts, including much of Cicero's *De republica,* from newly discovered palimpsests. A.I. Bekker, as well as editing many unknown Greek texts in the Paris collections (see above), was able, by use of newly discovered earlier and better manuscripts, to produce better editions of the classical authors than those then current. But the formulation of a technique of systematic recension (*i.e.,* analysis and evaluation of a manuscript tradition) was gradual, with its roots in the 18th century. Such New Testament scholars as J.A. Bengel (1687–1752) had established the principle that the witnesses to a text must be classified and their testimony evaluated according to their textual genealogy. For a time, the perceived barrier between "sacred" and "profane" texts limited the influence of such work on the analysis of "pagan" sources. During the first half of the 19th century the combined efforts of several scholars, notably Ritschl, Jakob Bernays (1824–81), and the Danish Latinist J.N. Madvig (1804–86), evolved the critical method usually associated with the name of Karl Lachmann (1793–1851) because it is most strikingly exemplified in his edition of Lucretius (1850).

The respect felt for Lachmann by such men as his friend and pupil Moritz Haupt turned into something like critical orthodoxy, however: the new techniques were rigorously applied by less gifted scholars, so that in this department of scholarship some work came to be distinguished by a blind confidence in the so-called scientific method as needing little intelligence in its handling. Madvig had realized the importance, restressed in the mid-20th cen-

tury, of allowing for inequalities and anomalies in an author's style; but these warnings were lost on those who, exuberantly confident in their own powers, proceeded to wholesale athetesis, or rejection of works as spurious, based on inconsistencies within a text. It was a similarly rigid insistence on analogical methods of criticism that marred the achievements of even such a great critic as the Dutch C.G. Cobet (1813–89) and so set a bad example to lesser scholars.

Developments in the study of ancient history and philosophy. Corresponding progress was made in the field of ancient history. Berthold Niebuhr, the pioneer in historical source criticism, applied a rational skepticism to ancient legends and traditions; he also promoted the collection of Latin inscriptions. J.G. Droysen (1808–84) wrote notable histories of Alexander the Great and of the Hellenistic Age; in fact, the very concept of a Hellenistic Age was his invention. Theodor Mommsen (1817–1903), starting as a professor of Roman law, made vast contributions to almost every branch of Roman studies, but particularly to the history of law and government and of the administration of the Roman provinces. He took a central part in the systematic collection of Latin inscriptions and was familiar with virtually every text and document relevant to Roman history. Mommsen's method of studying an entire civilization had influence on historical studies in general far beyond the limits of ancient history; perhaps the most distinguished of his pupils was the great sociologist Max Weber (1864–1920). In the second half of the century, Eduard Meyer (1855–1930), equipped with wide knowledge of Oriental as well as Greek and Latin sources, wrote an important history of the ancient world. Notable histories of Greece were brought out by Georg Busolt (1850–1920) and Karl Julius Beloch (1854–1929).

Greek philosophy

At the beginning of the 19th century Friedrich Schleiermacher (1768–1834), another of the scholars who gave the University of Berlin its special lustre, revitalized the study of Plato. Eduard Zeller (1814–1908) wrote a history of ancient philosophy that has been several times revised and is still useful. Later Hermann Diels (1848–1922) collected the fragments of pre-Socratic philosophers and of the so-called doxographers who preserved much of the evidence for our knowledge of ancient philosophy. The texts relevant to Epicureanism were edited by Hermann Usener (1834–1905), who employed the new methodology of comparative religion to throw much light on the religion of Greece, not disdaining the study of popular culture and of folklore as well; his work was continued by a line of pupils, and he had an important influence on the great art historian Aby Warburg (1866–1929). Late in the century Erwin Rohde (1845–98) wrote *Psyche,* an important study of Greek beliefs about the soul.

Nations other than Germany made more modest contributions to scholarship, most of them being more concerned with teaching than with research. (For the Italian contribution to papyrology, see below.) The literary scholarship of the French, though elegant and polished, was superficial in comparison with that of the Germans. It is significant that the leading Hellenist of France was the German-born Henri Weil (1818–1909). Great figures were exceptional; among them were J.A. Letronne (1787–1848), an archaeologist and epigraphist, and Paul-Émile Littré (1801–81), the famous lexicographer and positivist philosopher whose remarkable translation of Hippocrates emanated merely as a side interest of his philosophical vocation.

In England a notable contribution to ancient history and to the study of Plato and Aristotle came from the banker George Grote (1794–1871), who saw antiquity from the viewpoint of modern liberalism and utilitarianism. Later in the century the ancient universities produced a few distinguished scholars: H.A.J. Munro (1819–85) did important work on Catullus and Lucretius; Sir Richard Jebb (1841–1905) wrote a good commentary on the newly discovered Bacchylides and also one on Sophocles, which despite some technical deficiencies is still useful because of the author's rare feeling for Greek; and Ingram Bywater (1840–1914) contributed significantly to the study of Aristotle's *Ethics* and *Poetics* and acquired rare knowledge of the history of scholarship.

In the United States the leading figure was B.L. Gildersleeve (1831–1924), who insisted that American classical scholars should aim at the highest European standards. Germany was rightly taken as a model, and valuable work was done, especially in grammar, syntax, and linguistics, by such scholars as W.W. Goodwin (1831–1912), J.W. White (1849–1917), and H.W. Smyth (1857–1937). But too often it was not the admirable qualities of the best German scholars but the dryness, pedantry, and verbosity of the worst that were reproduced. This led to a reaction that went too far in the opposite direction and so did considerable damage. In archaeology, however, the vast resources of America were applied with ever-increasing effectiveness.

Developments in archaeology and art history. The foundation of the Instituto di Correspondenza Archeologica in Rome in 1829 provided an international centre for archaeological studies in Italy, which now progressed rapidly. Eduard Gerhard (1795–1867) founded the study of Greek vase painting as a scientific discipline; his report on the numerous Greek vases excavated from the Etruscan necropolis of Vulci (1831) was epoch-making. In Bonn, Welcker built up the first large collection of plaster casts of Greek sculpture. Another pioneer of the study of Greek art was his colleague Otto Jahn (1813–69), also an excellent Latinist. After the establishment of the Greek kingdom in 1830 the various European nations set up schools in Athens as they had done in Rome, and excavations on a large scale took place not only in Greece but all over the eastern Mediterranean world.

In archaeology the great impetus came from an amateur, Heinrich Schliemann (1822–90), whom no one can deprive of the credit for having guessed that remarkable finds might be made at Troy, Mycenae, and Tiryns, for having deliberately made a fortune so that he might do so, and for having discovered and promoted the great archaeologist Wilhelm Dörpfeld (1853–1940). In 1900 the ancient city of Knossos on Crete was excavated by the English archaeologist Sir Arthur Evans (1851–1941), which enabled the study of Mycenaean civilization to be supplemented by that of Minoan. The French excavated the two great Apollonian shrines at Delos and Delphi.

Papyri had been found in large numbers in the Epicurean library at Herculaneum discovered during the 18th century, and from 1878, when a roll turned up in Egypt, sporadic finds were made. From about the 1870s systematic excavation led to a steady stream of discoveries, mostly from al-Fayyūm, where the sunshine acts upon the soil in such a way as to preserve papyrus. The Italian Amedeo Peyron (1785–1870) was a pioneer in the new discipline of papyrology, as was Domenico Comparetti (1835–1927), the author of a famous book about the fortune of Virgil's works during the Middle Ages. The eminent Italian legal scholar and paleographer Girolamo Vitelli (1849–1935) became an expert papyrologist and had great personal influence. In Germany important papyri were published under the supervision of Wilhelm Schubart with the help of Wilamowitz-Moellendorff. In 1891 the *Constitution of Athens* by Aristotle and the poems of Herodas were published from a papyrus in the British Museum, and in 1897 they were followed by the poems of Bacchylides from the 5th century BC. In 1898 the Oxford scholars B.P. Grenfell and A.S. Hunt brought out the first volume of the series, still not concluded, that contains the texts of the papyri found by them at Oxyrhynchus. Documentary papyri supply useful evidence for law and government in Roman Egypt, and literary papyri supply a priceless supplement to the knowledge of Greek (and occasionally Latin) literature.

The rise of papyrology

The 19th century saw the beginning of many great enterprises, both individual and collective, that have equipped scholars with invaluable tools: collections of fragments, inscriptions, and works of art; and improved dictionaries, special lexica, handbooks, encyclopaedias, and catalogs of manuscripts. The invention of photography made it possible to produce facsimiles of manuscripts and documents and to distribute better likenesses of monuments and works of art. Many of these projects were sponsored by the various national academies, which were now linked by

the Association des Académies, the driving force of which was Mommsen.

The rise of professionalism. Associated with Germany was the movement toward what may be called professionalism during the second half of the 19th century. Though Wolf's example in founding a classical periodical in the vernacular had been followed elsewhere (*e.g.*, the English *Classical Journal,* 1810–29), journals written primarily by professional scholars for professional scholars only began to proliferate after about 1850. Coupled with this proliferation were the increased importance of universities, seminars, and academies (with their published proceedings) and the growing habit of early publication of, for instance, the Ph.D. dissertation, the academic "program," and the technical monograph.

Professional journals

Specialization was accompanied by a rise in technical standards of argument and presentation and a tendency toward the use of learned jargon—a phenomenon particularly noticeable in classical studies because of the contrast with earlier scholarly literature. An allied change was the replacement of Latin by the vernacular as a medium of scholarly intercourse and publication (with traditional exceptions, such as the preface and apparatus of a critical text). Thus, after about 1850 a classical scholar who wished to keep abreast of developments in his subject had to be able to read at the least English, French, German, Italian, and, in some cases, Russian. These changes had more immediate results in continental Europe and the United States; in England their effects were delayed in part by the insularity that characterized English scholarship after Bentley, in part by the concentration of the older universities on teaching, and a consequent distrust by tutors of a strong professoriate and of "pure research."

Late 19th-century developments in German scholarship. Germany made so vast a contribution to 19th-century classical scholarship that it would be impossible to name all of the eminent scholars of the period. But from a time rather earlier than the establishment of the German Empire (1871), signs of decline might be observed; the new methods had begun to harden into orthodoxy, mechanically applied by a mass of inferior practitioners. There was a strong tendency toward excessive emendation and deletion, and the overconscientious accumulation of details led to much dullness. From this situation German scholarship was to make a remarkable, though not complete, recovery, thanks to the generation of Ulrich von Wilamowitz-Moellendorff (1848–1931), who broke down the barriers that had grown up between the divisions of his subject, making important contributions to them all. He was the author of the first commentary on a Greek poem in which the entire apparatus of modern scholarship, encompassing not only literary knowledge but also that of history, art, archaeology, linguistics, and religion, was brought to bear on the elucidation of the work in question; this was his commentary on the *Herakles* of Euripides (1st edition, with a remarkable introduction to Attic tragedy, 1889; 2nd edition, 1895). Wilamowitz-Moellendorff produced many more texts and commentaries, besides important work on Greek history, religion, metre, and the history of scholarship. As a professor in Greifswald, Göttingen, and finally Berlin, he exercised a powerful influence.

Wilamowitz-Moellendorff

At the same time Eduard Schwartz (1851–1940) did much not only for the study of Greek history and literature but also for the history of the Christian Church; Georg Kaibel (1850–1901) advanced the study of Greek drama and of verse inscriptions; and Carl Robert (1850–1922) combined archaeological with literary expertise in remarkable fashion. Friedrich Leo (1851–1914) contributed significantly to Plautine studies and began a history of Latin literature of high quality. Jacob Wackernagel (1853–1938) of Basel and Wilhelm Schulze (1863–1935) used their mastery of comparative linguistics to throw light on Greek and Latin texts. Richard Reitzenstein (1861–1931) was eminent not only in the field of Greek literature and lexicography but also in that of ancient religion. Ludwig Traube (1861–1907) did important work in Latin paleography.

Classical scholarship in the 20th century. World War I dealt a heavy blow to classical studies, as to all humane letters, and the numbers of those studying Greek and Latin were noticeably affected; but scholars showed courage and energy in adapting themselves to new conditions. Wilamowitz-Moellendorff continued to be active, and his last decade saw more abundant and more important publications than any other of his career. His pupils produced much important detailed work: Felix Jacoby (1876–1959) began and carried far a learned edition of the fragments of the Greek historians; Paul Maas (1880–1964) showed rare expertise in Greek metrics, textual criticism, and paleography; Eduard Fraenkel (1888–1970) did valuable work on Plautus' relation to his Greek originals and later devoted to Aeschylus' *Agamemnon* one of the most learned of all commentaries; and Rudolf Pfeiffer (1889–1979) wrote a masterly commentary on Callimachus and an important history of classical scholarship.

Reacting against the classicism of the age of Goethe, scholars of the late 19th century saw the study of antiquity mainly from a historical standpoint: they accumulated masses of detail, which sometimes led to dryness, and tended to think exclusively in terms of concrete fact. Discontent arose with the recognition that an excessive preoccupation with the details of their development can harm the understanding of works of literature and thought. Attempts were made to revive classical scholarship by rescuing it from the domination of historical study. Werner Jaeger (1888–1961), an Aristotelian scholar who succeeded Wilamowitz-Moellendorff in his Berlin chair, attempted, without much success, to achieve this by institutional means. More was accomplished by Karl Reinhardt (1886–1958), who, though a devoted pupil of Wilamowitz-Moellendorff, had been in contact from his youth with the ideas of Nietzsche and of the circle around the poet Stefan George. Combining deep learning with refined sensibility, Reinhardt did important work on pre-Socratic philosophy and on Poseidonius and later on Sophocles, Aeschylus, and Homer.

Even before the start of World War II, National Socialist persecution had gravely damaged scholarship in Germany, the main centre of classical studies. The United States and, to an even greater extent, England benefited from the efflux of scholars from the Continent. Jaeger and two other eminent pupils of Wilamowitz-Moellendorff, Paul Friedländer (1882–1968) and Hermann Fränkel (1888–1977), spent the rest of their lives in the United States. So did the Russian M.I. Rostovtzev (1870–1952), who made a vast contribution to the study of the social and economic history of the ancient world. Thaddeus Zielinski (1859–1944), the Polish scholar who did important work on Ciceronian clausulae (clauses) and other topics, was murdered by the Nazis. Eduard Norden (1868–1941), who studied the formal prose of the ancients and did important work on ancient religion and on Latin literature, died in Switzerland. Jacoby, Maas, Fraenkel, and Pfeiffer, as well as the eminent archaeologist Paul Jacobsthal (1880–1957), settled in England, where Fraenkel in particular taught most effectively, creating links between English and continental scholarship. Pfeiffer, like Kurt von Fritz (1900–85), who spent the war years in America, returned to Germany.

Impact of Nazism

In Italy the school founded by Vitelli continued under the leadership of Giorgio Pasquali (1887–1952), a pupil of Schwartz and Leo, and Gaetano de Sanctis (1870–1957) did important work on ancient history. In Sweden Einar Löfstedt (1880–1955) and his school threw much light on Vulgar Latin and indirectly on Latin in general, and M.P. Nilsson (1874–1967) wrote a learned history of Greek religion.

In France classical studies to some degree slumbered under the conservative establishment, but Antoine Meillet (1866–1936) and others advanced the study of linguistics, and Louis Gernet (1882–1962) founded an important school of scholars who applied the techniques of modern sociology and anthropology to the study of antiquity.

In England A.E. Housman (1859–1936) continued with great distinction the tradition of exclusively textual study, editing Juvenal, Lucan, and most notably Manilius. J.D. Denniston (1887–1949) made a valuable study of the Greek particles. Edgar Lobel (1887–1981) from 1927 edited the literary papyri from Oxyrhynchus with unri-

valed expertise. Sir Denys Page (1908–78) edited many Greek poetical texts with great success. Gilbert Murray (1866–1957) was not only a literary scholar but, like Jane Harrison (1850–1928), a pioneer in the use of anthropological and sociological methods in the study of antiquity. F.M. Cornford (1874–1943) shared this interest but went on to contribute significantly to the study of Plato and the pre-Socratics. E.R. Dodds, starting with Neoplatonism, applied psychological as well as anthropological knowledge to the study of early Greek thought, also writing excellent commentaries on Euripides' *Bacchae* and Plato's *Gorgias*. Sir John Beazley (1885–1970), with deep learning and refined sensibility, put the whole study of Greek vase painting on a new basis by applying the method of the 19th-century Italian art critic Giovanni Morelli to the identification of individual painters.

The Homeric Question

The way in which research may (and indeed must) transcend the conventional limits of individual disciplines is exemplified during this period in the history of the Homeric Question: the efforts of scholars in such diverse fields as linguistics, archaeology, Hittite studies, folklore, and comparative oral literature have materially advanced understanding of the poems. The problem was transformed by the proof of an American scholar, Milman Parry (1902–35), that the poems are typical of a poetic tradition that has passed through a long phase of oral transmission.

Excavation continued, despite many political and financial difficulties, and a steady stream of discoveries came from Greece, Italy, and other Mediterranean lands. Perhaps the most exciting new find after World War II was the discovery by the Greek archaeologist Spyridon Marinatos of a Minoan town, with fine and well-preserved frescoes, on the island of Thera. Although large-scale excavations in search of papyri have been discontinued for many years, new papyri have not ceased to be discovered. Since World War II the authors who have benefited most have been Callimachus, Menander, and Stesichorus. In 1952 Michael Ventris showed that the language of the so-called Linear B syllabic script on clay tablets found at Mycenae and other places is Greek, thus throwing light on a far earlier stage of the language than had previously been known.

Linear B script

The history of classical scholarship has continued to be one of activity and progress. The publication of new inscriptions and of new papyri and other manuscripts has yielded important new material, and, considering the limited resources available, the task of presenting the texts of literary works and documents in up-to-date editions has been carried out with considerable success. Lately the Hellenistic and Imperial periods have received greater emphasis and have been given greater credit for their achievements.

But such are the threats presented by social change and utilitarian pressures that heroic efforts will be needed if progress is to continue. In Europe at the beginning of the 20th century many schools gave a good grounding in the ancient languages. This is now no longer the case and, as a result, the years when the memory is at its best for learning new languages are wasted. In the United States, vast reserves not only of money but also of talent and enthusiasm make a large contribution to classical studies, but progress is impeded not only by the failure of the schools to teach the ancient languages but also by the materialism and utilitarianism that are gaining ground both there and in Europe.

BIBLIOGRAPHY. JOHN EDWIN SANDYS, *A History of Classical Scholarship*, 3 vol. (1903–08, reissued 1967), while not a critical study, is useful for factual information. RUDOLF PFEIFFER, *History of Classical Scholarship from the Beginnings to the End of the Hellenistic Age* (1968), is a masterly critical survey, and *History of Classical Scholarship from 1300 to 1850* (1976), contains much valuable material but is uneven and lacks adequate treatment of the important 19th-century period. The best brief survey is U. VON WILAMOWITZ-MOELLENDORFF, *History of Classical Scholarship* (1982; originally published in German, 1921). A history of classical scholarship in antiquity is found in JAMES E.G. ZETZEL, *Latin Textual Criticism in Antiquity* (1981, reprinted 1984). For a discussion of the transmission of Greek and Latin literature, see L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars*, 2nd rev. ed. (1974); and L.D. REYNOLDS (ed.), *Texts and Transmission: A Survey of the Latin Classics* (1983). N.G. WILSON, *Scholars of Byzantium* (1983), chronicles the history of Byzantine scholarship. ROBERTO WEISS, *Medieval and Humanist Greek* (1977), is a collection of essays detailing the use of the Greek language in the Latin Middle Ages. Weiss also covers a later age in *The Renaissance Discovery of Classical Antiquity* (1969, reissued 1973). ANTHONY GRAFTON, *Joseph Scaliger: A Study in the History of Classical Scholarship*, vol. 1, *Textual Criticism and Exegesis* (1983), presents a biography of the 16th-century French classicist. For English scholarship, see C.O. BRINK, *English Classical Scholarship: Historical Reflections on Bentley, Porson and Housman* (1985); and M.L. CLARKE, *Greek Studies in England, 1700–1830* (1945). On the history of Greek vase painting, see R.M. COOK, *Greek Painted Pottery*, 2nd ed. (1972). On the history of papyrology, see E.G. TURNER, *Greek Papyri: An Introduction* (1968, reissued 1980). E.J. KENNEY, *The Classical Text: Aspects of Editing in the Age of the Printed Book* (1974); and SEBASTIANO TIMPANARO, *La genesi del metodo del Lachmann*, new rev. ed. (1981), treat the development of textual criticism. For a discussion of classical influences in the 19th and 20th centuries, see HUGH LLOYD-JONES, *Blood for the Ghosts* (1983), and *Classical Survivals: The Classics in the Modern World* (1982). In general see the collections of essays by ARNALDO MOMIGLIANO, many in English: vol. 1 appeared as *Contributo alla storia degli studi classici* (1955, reprinted 1979), and the most recent addition appeared as vol. 7, *Settimo contributo alla storia degli studi classici e del mondo antico* (1984).

(H.L.-J.)

The History of Science

On the simplest level, science is knowledge of the world of nature. There are many regularities in nature that mankind has had to recognize for survival since the emergence of *Homo sapiens* as a species. The Sun and the Moon periodically repeat their movements. Some motions, like the daily "motion" of the Sun, are simple to observe; others, like the annual "motion" of the Sun, are far more difficult. Both motions correlate with important terrestrial events. Day and night provide the basic rhythm of human existence; the seasons determine the migration of animals upon which humans depended for millennia for survival. With the invention of agriculture, the seasons became even more crucial, for failure to recognize the proper time for planting could lead to starvation. Science defined simply as knowledge of natural processes is universal among mankind, and it has existed since the dawn of human existence.

The mere recognition of regularities does not exhaust the full meaning of science, however. In the first place, regularities may be simply constructs of the human mind. Humans leap to conclusions; the mind cannot tolerate chaos, so it constructs regularities even when none objectively exists. Thus, for example, one of the astronomical "laws" of the Middle Ages was that the appearance of comets presaged a great upheaval, as the Norman Conquest of Britain followed the comet of 1066. True regularities must be established by detached examination of data. Science, therefore, must employ a certain degree of skepticism to prevent premature generalization.

Regularities, even when expressed mathematically as laws of nature, are not fully satisfactory to everyone. Some insist that genuine understanding demands explanations of the causes of the laws, but it is in the realm of causation that there is the greatest disagreement. Modern quantum mechanics, for example, has given up the quest for causation and today rests only on mathematical description. Modern biology, on the other hand, thrives on causal chains that permit the understanding of physiological and evolutionary processes in terms of the physical activities of entities such as molecules, cells, and organisms. But even if causation and explanation are admitted as necessary, there is little agreement on the kinds of causes that are permissible, or possible, in science. If the history of science is to make any sense whatsoever, it is necessary to deal with the past on its own terms, and the fact is that for most of the history of science natural philosophers appealed to causes that would be summarily rejected by modern scientists. Spiritual and divine forces were accepted as both real and necessary until the end of the 18th century and, in areas such as biology, deep into the 19th century as well.

Certain conventions governed the appeal to God or the gods or to spirits. Gods and spirits, it was held, could not be completely arbitrary in their actions; otherwise the proper response would be propitiation, not rational investigation. But since the deity or deities were themselves rational, or bound by rational principles, it was possible for humans to uncover the rational order of the world. Faith in the ultimate rationality of the creator or governor of the world could actually stimulate original scientific work. Kepler's laws, Newton's absolute space, and Einstein's rejection of the probabilistic nature of quantum mechanics were all based on theological, not scientific, assumptions. For sensitive interpreters of phenomena, the ultimate intelligibility of nature has seemed to demand some rational guiding spirit. A notable expression of this idea is Einstein's statement that the wonder is not that mankind comprehends the world, but that the world is comprehensible.

Science, then, is to be considered in this article as knowledge of natural regularities that is subjected to some degree of skeptical rigour and explained by rational causes. One final caution is necessary. Nature is known only through the senses, of which sight, touch, and hearing are the dominant ones, and the human notion of reality is skewed toward the objects of these senses. The invention of such instruments as the telescope, the microscope, and the Geiger counter has brought an ever-increasing range of phenomena within the scope of the senses. Thus, scientific knowledge of the world is only partial, and the progress of science follows the ability of humans to make phenomena perceivable.

This article provides a broad survey of the development of science as a way of studying and understanding the world, from the primitive stage of noting important regularities in nature to the epochal revolution in our notion of what constitutes reality that has occurred in 20th-century physics. More detailed treatments of the histories of specific sciences, including developments of the later 20th century, may be found in the articles BIOLOGICAL SCIENCES; EARTH SCIENCES; and PHYSICAL SCIENCES. See also the references in the *Propædia,* Part Ten, Division III.

The article is divided into the following sections:

Science as natural philosophy 32
- Precritical science 32
  - China 33
  - India 33
  - America 33
  - The Middle East 33
- Greek science 34
  - The birth of natural philosophy 34
  - Aristotle and Archimedes 34
  - Medicine 35
  - Science in Rome and Christianity 35
- Science in Islām 35
- Medieval European science 36

The rise of modern science 36
- The authority of phenomena 36
- The scientific revolution 37
  - Copernicus 37
  - Tycho, Kepler, and Galileo 37
  - Newton 38
  - The diffusion of scientific method 38
- The classic age of science 39
  - Mechanics 39
  - Chemistry 39
  - The imponderable fluids 39
  - Science and the Industrial Revolution 39
  - The Romantic revolt 40
  - The founding of modern biology 40
- The 20th-century revolution 41

Bibliography 41

Science as natural philosophy

PRECRITICAL SCIENCE

Science, as it has been defined above, made its appearance before writing. It is necessary, therefore, to infer from archaeological remains what was the content of that science. From cave paintings and from apparently regular scratches on bone and reindeer horn, it is known that prehistoric humans were close observers of nature who carefully tracked the seasons and times of the year. About 2500 BC there was a sudden burst of activity that seems to have had clear scientific importance. Great Britain and northwest-

ern Europe contain large stone structures from that era, the most famous of which is Stonehenge on the Salisbury Plain in England, that are remarkable from a scientific point of view. Not only do they reveal technical and social skills of a high order—it was no mean feat to move such enormous blocks of stone considerable distances and place them in position—but the basic conception of Stonehenge and the other megalithic structures also seems to combine religious and astronomical purposes. Their layouts suggest a degree of mathematical sophistication that was first suspected only in the mid-20th century. Stonehenge is a circle, but some of the other megalithic structures are egg-shaped and, apparently, constructed on mathematical principles that require at least practical knowledge of the Pythagorean theorem that the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides. This theorem, or at least the Pythagorean numbers that can be generated by it, seems to have been known throughout Asia, the Middle East, and Neolithic Europe two millennia before the birth of Pythagoras.

This combination of religion and astronomy was fundamental to the early history of science. It is found in Mesopotamia, Egypt, China (although to a much lesser extent than elsewhere), Central America, and India. The spectacle of the heavens, with the clearly discernible order and regularity of most heavenly bodies highlighted by extraordinary events such as comets and novae and the peculiar motions of the planets, obviously was an irresistible intellectual puzzle to early mankind. In its search for order and regularity, the human mind could do no better than to seize upon the heavens as the paradigm of certain knowledge. Astronomy was to remain the queen of the sciences (welded solidly to theology) for the next 4,000 years.

Science, in its mature form, developed only in the West. But it is instructive to survey the protoscience that appeared in other areas, especially in light of the fact that until quite recently this knowledge was often, as in China, far superior to Western science.

China. As has already been noted, astronomy seems everywhere to have been the first science to emerge. Its intimate relation to religion gave it a ritual dimension that then stimulated the growth of mathematics. Chinese savants, for example, early devised a calendar and methods of plotting the positions of stellar constellations. Since changes in the heavens presaged important changes on the Earth (for the Chinese considered the universe to be a vast organism in which all elements were connected), astronomy and astrology were incorporated into the system of government from the very dawn of the Chinese state in the 2nd millennium BC. As the Chinese bureaucracy developed, an accurate calendar became absolutely necessary to the maintenance of order. The result was a system of astronomical observations and records unparalleled elsewhere, thanks to which there are, today, star catalogs and observations of eclipses and novae that go back for millennia.

The Chinese calendar

In other sciences, too, the overriding emphasis was on practicality, for the Chinese, almost alone among ancient peoples, did not fill the cosmos with gods and demons whose arbitrary wills determined events. Order was inherent and, therefore, expected. It was for man to detect and describe this order and to profit from it. Chemistry (or, rather, alchemy), medicine, geology, geography, and technology were all encouraged by the state and flourished. Practical knowledge of a high order permitted the Chinese to deal with practical problems for centuries on a level not attained in the West until the Renaissance.

India. Far less is known about science in India, largely because few scholars have investigated it. It is known that astronomy was studied for calendrical purposes to set the times for both practical and religious tasks. Primary emphasis was placed on solar and lunar motions, the fixed stars serving only as a background against which these luminaries moved. Indian mathematics, on the other hand, seems to have been quite advanced, with particular sophistication in geometrical and algebraic techniques. This latter branch was undoubtedly stimulated by the flexibility of the Indian system of numeration that later was to come into the West as the Hindu-Arabic numerals. Indian thought, however, was primarily philosophical and otherworldly and was concerned more with escaping this world than with understanding it.

America. Quite independently of China, India, and the other civilizations of Europe and Asia, the Maya of Central America, building upon older cultures, created a complex society in which astronomy and astrology played important roles. Determination of the calendar, again, had both practical and religious significance. Solar and lunar eclipses were important, as was the position of the bright planet Venus. No sophisticated mathematics are known to have been associated with this astronomy, but the Mayan calendar was both ingenious and the result of careful observation.

The Middle East. In the cradles of Western civilization in Egypt and Mesopotamia, there were two rather different situations. In Egypt, as in China, there was an assumption of cosmic order guaranteed by a host of benevolent gods. But unlike China, whose rugged geography often produced disastrous floods, earthquakes, and violent storms that destroyed crops, Egypt was surpassingly placid and delightful. Life was, in fact, so pleasant that the major concern of most Egyptians was over leaving it. Egyptians found it difficult to believe that all ended with death; enormous intellectual and physical labour, therefore, was devoted to preserving life after death. Both Egyptian theology and the pyramids are testaments to this preoccupation. Science did not flourish in this atmosphere. All of the important questions were answered by religion, so the Egyptians did not concern themselves overmuch with speculations about the universe. The stars and the planets had astrological significance in that the major heavenly bodies were assumed to "rule" the land when they were in the ascendant (from the succession of these "rules" came the seven-day week, after the five planets and the Sun and the Moon), but astronomy was largely limited to the calendrical calculations necessary to predict the annual life-giving flood of the Nile. None of this required much mathematics, and there was, consequently, little of any importance.

Mesopotamia was more like China. The life of the land depended upon the two great rivers, the Tigris and the Euphrates, as that of China depended upon the Huang Ho (Yellow River) and the Yangtze. The land was harsh and made habitable only by extensive damming and irrigation works. Storms, insects, floods, and invaders made life insecure. To create a stable society required both great technological skill, for the creation of hydraulic works, and the ability to hold off the forces of disruption. These latter were early identified with powerful and arbitrary gods who dominated Mesopotamian theology. The cities of the plain were centred on temples run by a priestly caste whose functions included the planning of major public works, like canals, dams, and irrigation systems, the allocation of the resources of the city to its members, and the averting of a divine wrath that could wipe everything out.

Public works in Mesopotamia

Mathematics and astronomy thrived under these conditions. The number system, probably drawn from the system of weights and coinage, was based on 60 (it was in ancient Mesopotamia that the system of degrees, minutes, and seconds developed) and was adapted to a practical arithmetic. The heavens were the abode of the gods, and because heavenly phenomena were thought to presage terrestrial disasters, they were carefully observed and recorded. Out of these practices grew, first, a highly developed mathematics that went far beyond the requirements of daily business, and then, some centuries later, a descriptive astronomy that was the most sophisticated of the ancient world until the Greeks took it over and perfected it.

Nothing is known of the motives of these early mathematicians for carrying their studies beyond the calculations of volumes of dirt to be removed from canals and the provisions necessary for work parties. It may have been simply intellectual play—the role of playfulness in the history of science should not be underestimated—that led them onward to abstract algebra. There are texts from about 1700 BC that are remarkable for their mathematical suppleness. Babylonian mathematicians knew the

Pythagorean relationship well and used it constantly. They could solve simple quadratic equations and could even solve problems in compound interest involving exponents. From about a millennium later there are texts that utilize these skills to provide a very elaborate mathematical description of astronomical phenomena.

Although China and Mesopotamia provide examples of exact observation and precise description of nature, what is missing is explanation in the scientific mode. The Chinese assumed a cosmic order that was vaguely founded on the balance of opposite forces (yin–yang) and the harmony of the five elements (water, wood, metal, fire, and earth). Why this harmony obtained was not discussed. Similarly, the Egyptians found the world harmonious because the gods willed it so. For Babylonians and other Mesopotamian cultures, order existed only so long as all-powerful and capricious gods supported it. In all these societies, humans could describe nature and use it, but to understand it was the function of religion and magic, not reason. It was the Greeks who first sought to go beyond description and to arrive at reasonable explanations of natural phenomena that did not involve the arbitrary will of the gods. Gods might still play a role, as indeed they did for centuries to come, but even the gods were subject to rational laws.

GREEK SCIENCE

The birth of natural philosophy. There seems to be no good reason why the Hellenes, clustered in isolated city-states in a relatively poor and backward land, should have struck out into intellectual regions that were only dimly perceived, if at all, by the splendid civilizations of the Yangtze, the Tigris and Euphrates, and the Nile valleys. There were many differences between ancient Greece and the other civilizations, but perhaps the most significant was religion. What is striking about Greek religion, in contrast to the religions of Mesopotamia and Egypt, is its puerility. Both of the great river civilizations evolved complex theologies that served to answer most, if not all, of the large questions about mankind's place and destiny. Greek religion did not. It was, in fact, little more than a collection of folk tales, more appropriate to the campfire than to the temple. Perhaps this was the result of the collapse of an earlier Greek civilization, now called Mycenaean, toward the end of the 2nd millennium BC, when a dark age descended upon Greece that lasted for three centuries. All that was preserved were stories of gods and men, passed along by poets, that dimly reflected Mycenaean values and events. Such were the great poems of Homer, the *Iliad* and the *Odyssey,* in which heroes and gods mingled freely with one another. Indeed, they mingled too freely, for the gods appear in these tales as little more than immortal adolescents whose tricks and feats, when compared with the concerns of a Marduk or Jehovah, are infantile. There really was no Greek theology in the sense that theology provides a coherent and profound explanation of the workings of both the cosmos and the human heart. Hence, there were no easy answers to inquiring Greek minds. The result was that ample room was left for a more penetrating and ultimately more satisfying mode of inquiry. Thus were philosophy and its oldest offspring, science, born.

The absence of theology

The first natural philosopher, according to Hellenic tradition, was Thales of Miletus, who flourished in the 6th century BC. We know of him only through later accounts, for nothing he wrote has survived. He is supposed to have predicted a solar eclipse in 585 BC and to have invented the formal study of geometry in his demonstration of the bisecting of a circle by its diameter. Most importantly, he tried to explain all observed natural phenomena in terms of the changes of a single substance, water, which can be seen to exist in solid, liquid, and gaseous states. What for Thales guaranteed the regularity and rationality of the world was the innate divinity in all things that directed them to their divinely appointed ends. From these ideas there emerged two characteristics of classical Greek science. The first was the view of the universe as an ordered structure (the Greek *kósmos* means "order"). The second was the conviction that this order was not that of a mechanical contrivance but that of an organism; all parts of the universe had purposes in the overall scheme of things, and objects moved naturally toward the ends they were fated to serve. This motion toward ends is called teleology and, with but few exceptions, it permeated Greek as well as much later science.

Thales inadvertently made one other fundamental contribution to the development of natural science. By naming a specific substance as the basic element of all matter, Thales opened himself to criticism, which was not long in coming. His own disciple, Anaximander, was quick to argue that water could not be the basic substance. His argument was simple: water, if it is anything, is essentially wet; nothing can be its own contradiction. Hence, if Thales were correct, the opposite of wet could not exist in a substance, and that would preclude all of the dry things that are observed in the world. Therefore, Thales was wrong. Here was the birth of the critical tradition that is fundamental to the advance of science.

Thales' conjectures set off an intellectual explosion, most of which was devoted to increasingly refined criticisms of his doctrine of fundamental matter. Various single substances were proposed and then rejected, ultimately in favour of a multiplicity of elements that could account for such opposite qualities as wet and dry, hot and cold. Two centuries after Thales, most natural philosophers accepted a doctrine of four elements: earth (cold and dry), fire (hot and dry), water (cold and wet), and air (hot and wet). All bodies were made from these four.

The presence of the elements only guaranteed the presence of their qualities in various proportions. What was not accounted for was the form these elements took, which served to differentiate natural objects from one another. The problem of form was first attacked systematically by the philosopher and cult leader Pythagoras in the 6th century BC. Legend has it that Pythagoras became convinced of the primacy of number when he realized that the musical notes produced by a monochord were in simple ratio to the length of the string. Qualities (tones) were reduced to quantities (numbers in integral ratios). Thus was born mathematical physics, for this discovery provided the essential bridge between the world of physical experience and that of numerical relationships. Number provided the answer to the question of the origin of forms and qualities.

Pythagoras

Aristotle and Archimedes. Hellenic science was built upon the foundations laid by Thales and Pythagoras. It reached its zenith in the works of Aristotle and Archimedes. Aristotle represents the first tradition, that of qualitative forms and teleology. He was, himself, a biologist whose observations of marine organisms were unsurpassed until the 19th century. Biology is essentially teleological—the parts of a living organism are understood in terms of what they do in and for the organism—and Aristotle's biological works provided the framework for the science until the time of Charles Darwin. In physics, teleology is not so obvious, and Aristotle had to impose it on the cosmos. From Plato, his teacher, he inherited the theological proposition that the heavenly bodies (stars and planets) are literally divine and, as such, perfect. They could, therefore, move only in perfect, eternal, unchanging motion, which, by Plato's definition, meant perfect circles. The Earth, being obviously not divine, and inert, was at the centre. From the Earth to the sphere of the Moon, all things constantly changed, generating new forms and then decaying back into formlessness. Above the Moon the cosmos consisted of contiguous and concentric crystalline spheres moving on axes set at angles to one another (this accounted for the peculiar motions of the planets) and deriving their motion either from a fifth element that moved naturally in circles or from heavenly souls resident in the celestial bodies. The ultimate cause of all motion was a prime, or unmoved, mover (God) that stood outside the cosmos.

Aristotle was able to make a great deal of sense of observed nature by asking of any object or process: what is the material involved, what is its form and how did it get that form, and, most important of all, what is its purpose? What should be noted is that, for Aristotle, all activity that occurred spontaneously was natural. Hence, the proper means of investigation was observation. Experiment, that

is, altering natural conditions in order to throw light on the hidden properties and activities of objects, was unnatural and could not, therefore, be expected to reveal the essence of things. Experiment was thus not essential to Greek science.

The problem of purpose did not arise in the areas in which Archimedes made his most important contributions. He was, first of all, a brilliant mathematician whose work on conic sections and on the area of the circle prepared the way for the later invention of the calculus. It was in mathematical physics, however, that he made his greatest contributions to science. His mathematical demonstration of the law of the lever was as exact as a Euclidean proof in geometry. Similarly, his work on hydrostatics introduced and developed the method whereby physical characteristics, in this case specific gravity, which Archimedes discovered, are given mathematical shape and then manipulated by mathematical methods to yield mathematical conclusions that can be translated back into physical terms.

In one major area the Aristotelian and the Archimedean approaches were forced into a rather inconvenient marriage. Astronomy was the dominant physical science throughout antiquity, but it had never been successfully reduced to a coherent system. The Platonic-Aristotelian astral religion required that planetary orbits be circles. But, particularly after the conquests of Alexander the Great had made the observations and mathematical methods of the Babylonians available to the Greeks, astronomers found it impossible to reconcile theory and observation. Astronomy then split into two parts: one was physical and accepted Aristotelian theory in accounting for heavenly motion; the other ignored causation and concentrated solely on the creation of a mathematical model that could be used for computing planetary positions. Ptolemy, in the 2nd century AD, carried the latter tradition to its highest point in antiquity in his *Hē mathēmatikē syntaxis* ("The Mathematical Collection," better known under its Greek-Arabic title, *Almagest*).

Medicine. The Greeks not only made substantial progress in understanding the cosmos but also went far beyond their predecessors in their knowledge of the human body. Pre-Greek medicine had been almost entirely confined to religion and ritual. Disease was considered the result of divine disfavour and human sin, to be dealt with by spells, prayers, and other propitiatory measures. In the 5th century BC a revolutionary change came about that is associated with the name of Hippocrates. It was Hippocrates and his school who, influenced by the rise of natural philosophy, first insisted that disease was a natural, not a supernatural, phenomenon. Even maladies as striking as epilepsy, whose seizures appeared to be divinely caused, were held to originate in natural causes within the body.

Hippocrates

The height of medical science in antiquity was reached late in the Hellenistic period. Much work was done at the museum of Alexandria, a research institute set up under Greek influence in Egypt in the 3rd century BC to sponsor learning in general. The heart and the vascular system were investigated, as were the nerves and the brain. The organs of the thoracic cavity were described, and attempts were made to discover their functions. It was on these researches, and on his own dissections of apes and pigs, that the last great physician of antiquity, Galen of Pergamum, based his physiology. It was, essentially, a tripartite system in which so-called spirits—natural, vital, and animal—passed respectively through the veins, the arteries, and the nerves to vitalize the body as a whole. Galen's attempts to correlate therapeutics with his physiology were not successful, and so medical practice remained eclectic and a matter of the physician's choice. Usually the optimal choice was that propounded by the Hippocratics, who relied primarily on simple, clean living and the ability of the body to heal itself.

Science in Rome and Christianity. The apogee of Greek science in the works of Archimedes and Euclid coincided with the rise of Roman power in the Mediterranean. The Romans were deeply impressed by Greek art, literature, philosophy, and science, and after their conquest of Greece many Greek intellectuals served as household slaves tutoring noble Roman children. The Romans were a practical people, however, and, while they contemplated the Greek intellectual achievement with awe, they also could not help but ask what good it had done the Greeks. Roman common sense was what kept Rome great; science and philosophy were either ignored or relegated to rather low status. Even such a Hellenophile as the statesman and orator Cicero used Greek thought more to buttress the old Roman ways than as a source of new ideas and viewpoints.

The spirit of independent research was quite foreign to the Roman mind, so scientific innovation ground to a halt. The scientific legacy of Greece was condensed and corrupted into Roman encyclopaedias whose major function was entertainment rather than enlightenment. Typical of this spirit was the 1st-century-AD aristocrat Pliny the Elder, whose *Natural History* was a multivolume collection of myths, odd tales of wondrous creatures, magic, and some science, all mixed together uncritically for the titillation of other aristocrats. Aristotle would have been embarrassed by it.

At its height Rome incorporated a host of peoples with different customs, languages, and religions within its empire. One religious sect that proved more significant than the rest was Christianity. Jesus and his kingdom were not of this world, but his disciples and their followers were. This world could not be ignored, even though concern with worldly things could be dangerous to the soul. So the early Christians approached the worldly wisdom of their time with ambivalence: on the one hand, the rhetoric and the arguments of ancient philosophy were snares and delusions that might mislead the simple and the unwary; on the other hand, the sophisticated and the educated of the empire could not be converted unless the Christian message was presented in the terms and rhetoric of the philosophical schools. Before they knew it, the early Christians were enmeshed in metaphysical arguments, some of which involved physics. What, for example, was the nature of Jesus, in purely physical terms? How was it possible that anybody could have two different essential natures, as was claimed for Jesus? Such questions revealed how important knowledge of the arguments of Greek thinkers on the nature of substance could be to those engaged in founding a new theology.

Science and theology

Ancient learning, then, did not die with the fall of Rome and the occupation of the Western Empire by tribes of Germanic barbarians. To be sure, the lamp of learning burned very feebly, but it did not go out. Monks in monasteries faithfully copied out classics of ancient thought and early Christianity and preserved them for posterity. Monasteries continued to teach the elements of ancient learning, for little beyond the elementary survived in the Latin West. In the East, the Byzantine Empire remained strong, and there the ancient traditions continued. There was little original work done in the millennium following the fall of Rome, but the ancient texts were preserved along with knowledge of the ancient Greek language. This was to be a precious reservoir of learning for the Latin West in later centuries.

SCIENCE IN ISLĀM

The torch of ancient learning passed first to one of the invading groups that helped bring down the Eastern Empire. In the 7th century the Arabs, inspired by their new religion, burst out of the Arabian peninsula and laid the foundations of an Islāmic empire that eventually rivalled that of ancient Rome. To the Arabs, ancient science was a precious treasure. The Qur'ān, the sacred book of Islām, particularly praised medicine as an art close to God. Astronomy and astrology were believed to be one way of glimpsing what God willed for mankind. Contact with Hindu mathematics and the requirements of astronomy stimulated the study of numbers and of geometry. The writings of the Hellenes were, therefore, eagerly sought and translated, and thus much of the science of antiquity passed into Islāmic culture. Greek medicine, Greek astronomy and astrology, and Greek mathematics, together with the great philosophical works of Plato and, particularly, Aristotle, were assimilated in Islām by the end of the 9th century. Nor did the Arabs stop with assimilation.

They criticized and they innovated. Islāmic astronomy and astrology were aided by the construction of great astronomical observatories that provided accurate observations against which the Ptolemaic predictions could be checked. Numbers fascinated Islāmic thinkers, and this fascination served as the motivation for the creation of algebra (from Arabic *al-jabr*) and the study of algebraic functions.

MEDIEVAL EUROPEAN SCIENCE

Medieval Christendom confronted Islām chiefly in military crusades, in Spain and the Holy Land, and in theology. From this confrontation came the restoration of ancient learning to the West. The Reconquista in Spain gradually pushed the Moors south from the Pyrenees, and among the treasures left behind were Arabic translations of Greek works of science and philosophy. In 1085 the city of Toledo, with one of the finest libraries in Islām, fell to the Christians. Among the occupiers were Christian monks who quickly began the process of translating ancient works into Latin. By the end of the 12th century much of the ancient heritage was again available to the Latin West.

The medieval world was caricatured by thinkers of the 18th-century Enlightenment as a period of darkness, superstition, and hostility to science and learning. On the contrary, it was one of great technological vitality. The advances that were made may appear today as trifling, but that is because they were so fundamental. They included the horseshoe and the horse collar, without which horsepower cannot be efficiently exploited. The invention of the crank, the brace and bit, the wheelbarrow, and the flying buttress made possible the great Gothic cathedrals. Improvements in the gear trains of waterwheels and the development of windmills harnessed these sources of power with great efficiency. Mechanical ingenuity, building on experience with mills and power wheels, culminated in the 14th century in the mechanical clock, which not only set a new standard of chronometrical accuracy but also provided philosophers with a new metaphor for nature itself.

The mechanical clock

An equal amount of energy was devoted to achieving a scientific understanding of nature, but it is essential to understand to what use medieval thinkers put this kind of knowledge. As the fertility of the technology shows, medieval Europeans had no deep prejudices against utilitarian knowledge. But the areas in which scientific knowledge could find useful expression were few. Instead, science was viewed chiefly as a means of understanding God's creation and, thereby, the Godhead itself. The best example of this attitude is found in the medieval study of optics. Light, as Genesis makes clear, was among the first creations of God. The 12th–13th-century cleric-scholar Robert Grosseteste saw in light the first creative impulse. As light spread it created both space and matter, and, in its reflection from the outermost circle of the cosmos, it gradually solidified into the heavenly spheres. To understand the laws of the propagation of light was to understand, in some slight way, the nature of the creation.

In the course of studying light, particular problems were isolated and attacked. What, for example, is the rainbow? It is impossible to get close enough to a rainbow to see clearly what is going on, for as the observer moves, so too does the rainbow. It does seem to depend upon the presence of raindrops, so medieval investigators sought to bring the rainbow down from the skies into their studies. Insight into the nature of the rainbow could be achieved by simulating the conditions under which rainbows occur. For raindrops the investigators substituted hollow glass balls filled with water, so that the rainbow could be studied at leisure. Valid conclusions about rainbows could then be drawn by assuming the validity of the analogy between raindrops and water-filled globes. This involved the implicit assumptions that nature was simple (*i.e.*, governed by a few general laws) and that similar effects had similar causes. Such a nature was what could be expected of a rational, benevolent deity; hence, the assumption could be persuasively adopted.

Medieval philosophers were not content, as the above example shows, to repeat what the ancients had said. They subjected ancient texts to close critical scrutiny. Usually the intensity of the criticism was directly proportional to the theological significance of the problem involved. Such was the case with motion. Medieval philosophers examined all aspects of motion with great care, for the nature of motion had important theological implications. Thomas Aquinas used Aristotle's dictum, that everything that moves is moved by something else, to show that God must exist, for otherwise the existence of any motion would imply an infinite regression of prior causal motions.

It should be clear that there was no conscious conflict between science and religion in the Middle Ages. As Aquinas pointed out, God was the author of both the book of Scripture and the book of nature. The guide to nature was reason, the faculty that was the image of God in which mankind was made. Scripture was direct revelation, although it needed interpretation, for there were passages that were obscure or difficult. The two books, having the same author, could not contradict each other. For the short term, science and revelation marched hand in hand. Aquinas carefully wove knowledge of nature into his theology, as in his proof from motion of the existence of God. But if his scientific concepts of motion should ever be challenged, there would necessarily be a theological challenge as well. By working science into the very fabric of his theology, he virtually guaranteed that someday there would be conflict. Theologians would side with theology and scientists with science, to create a breach that neither particularly desired.

The glory of medieval science was its integration of science, philosophy, and theology into a magnificent and comprehensible whole. It can be best contemplated in the greatest of all medieval poems, Dante's *Divine Comedy*. Here was an essentially Aristotelian cosmos, finite and easily understood, over which God, his Son, and his saints reigned. Humanity and the Earth occupied the centre, as befitted their centrality in God's plan. The nine circles of hell were populated by humans whose exercise of their free will had led to their damnation. Purgatory contained lesser sinners still capable of salvation. The heavenly spheres were populated by the saved and the saintly. The natural hierarchy gave way to the spiritual hierarchy as one ascended toward the throne of God. Such a hierarchy was reflected in the social and political institutions of medieval Europe, and God, the supreme monarch, ruled his creation with justice and love. All fit together in a grand cosmic scheme, one not to be abandoned lightly.

Natural and spiritual hierarchy

The rise of modern science

THE AUTHORITY OF PHENOMENA

Even as Dante was writing his great work, deep forces were threatening the unitary cosmos he celebrated. The pace of technological innovation began to quicken. Particularly in Italy, the political demands of the time gave new importance to technology, and a new profession emerged, that of civil and military engineer. These people faced practical problems that demanded practical solutions. Leonardo da Vinci is certainly the most famous of them, though he was much more as well. A painter of genius, he closely studied human anatomy in order to give verisimilitude to his paintings. As a sculptor he mastered the difficult techniques of casting metal. As a producer-director of the form of Renaissance dramatic production called the masque, he devised complicated machinery to create special effects. But it was as a military engineer that he observed the path of a mortar bomb being lobbed over a city wall and insisted that the projectile did not follow two straight lines—a slanted ascent followed by a vertical drop—as Aristotle had said it must. Leonardo and his colleagues needed to know nature truly; no amount of book learning could substitute for actual experience, nor could books impose their authority upon phenomena. What Aristotle and his commentators asserted as philosophical necessity often did not gibe with what could be seen with one's own eyes. The hold of ancient philosophy was too strong to be broken lightly, but a healthy skepticism began to emerge.

The first really serious blow to the traditional acceptance of ancient authorities was the discovery of the New World at the end of the 15th century. Ptolemy, the great as-

tronomer and geographer, had insisted that only the three continents of Europe, Africa, and Asia could exist, and Christian scholars from St. Augustine on had accepted it, for otherwise men would have to walk upside down at the antipodes. But Ptolemy, St. Augustine, and a host of other authorities were wrong. The dramatic expansion of the known world also served to stimulate the study of mathematics, for wealth and fame awaited those who could turn navigation into a real and trustworthy science.

In large part the Renaissance was a time of feverish intellectual activity devoted to the complete recovery of the ancient heritage. To the Aristotelian texts that had been the foundation of medieval thought were added translations of Plato, with his vision of mathematical harmonies, of Galen, with his experiments in physiology and anatomy, and, perhaps most important of all, of Archimedes, who showed how theoretical physics could be done outside the traditional philosophical framework. The results were subversive.

The search for antiquity turned up a peculiar bundle of manuscripts that added a decisive impulse to the direction in which Renaissance science was moving. These manuscripts were taken to have been written by or to report almost at first hand the activities of the legendary priest, prophet, and sage Hermes Trismegistos. Hermes was supposedly a contemporary of Moses, and the Hermetic writings contained an alternative story of creation that gave man a far more prominent role than the traditional account. God had made man fully in his image: a creator, not just a rational animal. Man could imitate God by creating. To do so, he must learn nature's secrets, and this could be done only by forcing nature to yield them through the tortures of fire, distillation, and other alchemical manipulations. The reward for success would be eternal life and youth, as well as freedom from want and disease. It was a heady vision, and it gave rise to the notion that, through science and technology, man could bend nature to his wishes. This is essentially the modern view of science, and it should be emphasized that it occurs only in Western civilization. It is probably this attitude that permitted the West to surpass the East, after centuries of inferiority, in the exploitation of the physical world.

Hermes Trismegistos

The Hermetic tradition also had more specific effects. Inspired, as is now known, by late Platonist mysticism, the Hermetic writers had rhapsodized on enlightenment and on the source of light, the Sun. Marsilio Ficino, the 15th-century Florentine translator of both Plato and the Hermetic writings, composed a treatise on the Sun that came close to idolatry. A young Polish student visiting Italy at the turn of the 16th century was touched by this current. Back in Poland, he began to work on the problems posed by the Ptolemaic astronomical system. With the blessing of the church, which he served formally as a canon, Nicolaus Copernicus set out to modernize the astronomical apparatus by which the church made such important calculations as the proper dates for Easter and other festivals.

THE SCIENTIFIC REVOLUTION

Copernicus. In 1543, as he lay on his deathbed, Copernicus finished reading the proofs of his great work; he died just as it was published. His *De revolutionibus orbium coelestium* (*On the Revolutions of the Celestial Spheres*) was the opening shot in a revolution whose consequences were greater than those of any other intellectual event in the history of mankind. The scientific revolution radically altered the conditions of thought and of material existence in which the human race lives, and its effects are not yet exhausted.

All this was caused by Copernicus' daring in placing the Sun, not the Earth, at the centre of the cosmos. Copernicus actually cited Hermes Trismegistos to justify this idea, and his language was thoroughly Platonic. But he meant his work as a serious work in astronomy, not philosophy, so he set out to justify it observationally and mathematically. The results were impressive. At one stroke, Copernicus reduced a complexity verging on chaos to elegant simplicity. The apparent back-and-forth movements of the planets, which required prodigious ingenuity to accommodate within the Ptolemaic system, could be accounted for just in terms of the Earth's own orbital motion added to or subtracted from the motions of the planets. Variation in planetary brightness was also explained by this combination of motions. The fact that Mercury and Venus were never found opposite the Sun in the sky Copernicus explained by placing their orbits closer to the Sun than that of the Earth. Indeed, Copernicus was able to place the planets in order of their distances from the Sun by considering their speeds and thus to construct a system of the planets, something that had eluded Ptolemy. This system had a simplicity, coherence, and aesthetic charm that made it irresistible to those who felt that God was the supreme artist. His was not a rigorous argument, but aesthetic considerations are not to be ignored in the history of science.

Copernicus did not solve all of the difficulties of the Ptolemaic system. He had to keep some of the cumbrous apparatus of epicycles and other geometrical adjustments, as well as a few Aristotelian crystalline spheres. The result was neater, but not so striking that it commanded immediate universal assent. Moreover, there were some implications that caused considerable concern: Why should the crystalline orb containing the Earth circle the Sun? And how was it possible for the Earth itself to revolve on its axis once in 24 hours without hurling all objects, including humans, off its surface? No known physics could answer these questions, and the provision of such answers was to be the central concern of the scientific revolution.

More was at stake than physics and astronomy, for one of the implications of the Copernican system struck at the very foundations of contemporary society. If the Earth revolved around the Sun, then the apparent positions of the fixed stars should shift as the Earth moves in its orbit. Copernicus and his contemporaries could detect no such shift (called stellar parallax), and there were only two interpretations possible to explain this failure. Either the Earth was at the centre, in which case no parallax was to be expected, or the stars were so far away that the parallax was too small to be detected. Copernicus chose the latter and thereby had to accept an enormous cosmos consisting mostly of empty space. God, it had been assumed, did nothing in vain, so for what purposes might he have created a universe in which the Earth and mankind were lost in immense space? To accept Copernicus was to give up the Dantean cosmos. The Aristotelian hierarchy of social place, political position, and theological gradation would vanish, to be replaced by the flatness and plainness of Euclidean space. It was a grim prospect and not one that recommended itself to most 16th-century intellectuals, and so Copernicus' grand idea remained on the periphery of astronomical thought. All astronomers were aware of it, some measured their own views against it, but only a small handful eagerly accepted it.

In the century and a half following Copernicus, two easily discernible scientific movements developed. The first was critical, the second, innovative and synthetic. They worked together to bring the old cosmos into disrepute and, ultimately, to replace it with a new one. Although they existed side by side, their effects can more easily be seen if they are treated separately.

Tycho, Kepler, and Galileo. The critical tradition began with Copernicus. It led directly to the work of Tycho Brahe, who measured stellar and planetary positions more accurately than had anyone before him. But measurement alone could not decide between Copernicus and Ptolemy, and Tycho insisted that the Earth was motionless. Copernicus did persuade Tycho to move the centre of revolution of all other planets to the Sun. To do so, he had to abandon the Aristotelian crystalline spheres that otherwise would collide with one another. Tycho also cast doubt upon the Aristotelian doctrine of heavenly perfection, for when, in the 1570s, a comet and a new star appeared, Tycho showed that they were both above the sphere of the Moon. Perhaps the most serious critical blows struck were those delivered by Galileo after the invention of the telescope. In quick succession, he announced that there were mountains on the Moon, satellites circling Jupiter, and spots upon the Sun. Moreover, the Milky Way was

Invention of the telescope

composed of countless stars whose existence no one had suspected until Galileo saw them. Here was criticism that struck at the very roots of Aristotle's system of the world.

At the same time Galileo was searching the heavens with his telescope, in Germany Johannes Kepler was searching them with his mind. Tycho's precise observations permitted Kepler to discover that Mars (and, by analogy, all the other planets) did not revolve in a circle at all, but in an ellipse, with the Sun at one focus. Ellipses tied all the planets together in grand Copernican harmony. The Keplerian cosmos was most un-Aristotelian, but Kepler hid his discoveries by burying them in almost impenetrable Latin prose in a series of works that did not circulate widely.

What Galileo and Kepler could not provide, although they tried, was an alternative to Aristotle that made equal sense. If the Earth revolves on its axis, then why do objects not fly off it? And why do objects dropped from towers not fall to the west as the Earth rotates to the east beneath them? And how is it possible for the Earth, suspended in empty space, to go around the Sun—whether in circles or ellipses—without anything pushing it? The answers were long in coming.

Galileo attacked the problems of the Earth's rotation and its revolution by logical analysis. Bodies do not fly off the Earth because they are not really revolving rapidly, even though their speed is high. In revolutions per minute, any body on the Earth is going very slowly and, therefore, has little tendency to fly off. Bodies fall to the base of towers from which they are dropped because they share with the tower the rotation of the Earth. Hence, bodies already in motion preserve that motion when another motion is added. So, Galileo deduced, a ball dropped from the top of a mast of a moving ship would fall at the base of the mast. If the ball were allowed to move on a frictionless horizontal plane, it would continue to move forever. Hence, Galileo concluded, the planets, once set in circular motion, continue to move in circles forever. Therefore, Copernican orbits exist. Galileo never acknowledged Kepler's ellipses; to do so would have meant abandoning his solution to the Copernican problem.

Kepler realized that there was a real problem with planetary motion. He sought to solve it by appealing to the one force that appeared to be cosmic in nature, namely magnetism. The Earth had been shown to be a giant magnet by William Gilbert in 1600, and Kepler seized upon this fact. A magnetic force, Kepler argued, emanated from the Sun and pushed the planets around in their orbits, but he was never able to quantify this rather vague and unsatisfactory idea.

By the end of the first quarter of the 17th century Aristotelianism was rapidly dying, but there was no satisfactory system to take its place. The result was a mood of skepticism and unease, for, as one observer put it, "The new philosophy calls all in doubt." It was this void that accounted largely for the success of a rather crude system proposed by René Descartes. Matter and motion were taken by Descartes to explain everything by means of mechanical models of natural processes, even though he warned that such models were not the way nature probably worked. They provided merely "likely stories," which seemed better than no explanation at all.

Descartes

Armed with matter and motion, Descartes attacked the basic Copernican problems. Bodies once in motion, Descartes argued, remain in motion in a straight line unless and until they are deflected from this line by the impact of another body. All changes of motion are the result of such impacts. Hence, the ball falls at the foot of the mast because, unless struck by another body, it continues to move with the ship. Planets move around the Sun because they are swept around by whirlpools of a subtle matter filling all space. Similar models could be constructed to account for all phenomena; the Aristotelian system could be replaced by the Cartesian. There was one major problem, however, and it sufficed to bring down Cartesianism. Cartesian matter and motion had no purpose, nor did Descartes's philosophy seem to need the active participation of a deity. The Cartesian cosmos, as Voltaire later put it, was like a watch that had been wound up at the creation and continues ticking to eternity.

Newton. The 17th century was a time of intense religious feeling, and nowhere was that feeling more intense than in Great Britain. There a devout young man, Isaac Newton, was finally to discover the way to a new synthesis in which truth was revealed and God was preserved.

Newton was both an experimental and a mathematical genius, a combination that enabled him to establish both the Copernican system and a new mechanics. His method was simplicity itself: "from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena." Newton's genius guided him in the selection of phenomena to be investigated, and his creation of a fundamental mathematical tool—the calculus (simultaneously invented by Gottfried Leibniz)—permitted him to submit the forces he inferred to calculation. The result was *Philosophiae Naturalis Principia Mathematica* (*Mathematical Principles of Natural Philosophy,* usually called simply the *Principia*), which appeared in 1687. Here was a new physics that applied equally well to terrestrial and celestial bodies. Copernicus, Kepler, and Galileo were all justified by Newton's analysis of forces. Descartes was utterly routed.

Newton's three laws of motion and his principle of universal gravitation sufficed to regulate the new cosmos, but only, Newton believed, with the help of God. Gravity, he more than once hinted, was direct divine action, as were all forces for order and vitality. Absolute space, for Newton, was essential, because space was the "sensorium of God," and the divine abode must necessarily be the ultimate coordinate system. Finally, Newton's analysis of the mutual perturbations of the planets caused by their individual gravitational fields predicted the natural collapse of the solar system unless God acted to set things right again.

The role of the divine

The diffusion of scientific method. The publication of the *Principia* marks the culmination of the movement begun by Copernicus and, as such, has always stood as the symbol of the scientific revolution. There were, however, similar attempts to criticize, systematize, and organize natural knowledge that did not lead to such dramatic results. In the same year as Copernicus' great volume, there appeared an equally important book on anatomy: Andreas Vesalius' *De humani corporis fabrica* ("On the Fabric of the Human Body," called the *De fabrica*), a critical examination of Galen's anatomy in which Vesalius drew on his own studies to correct many of Galen's errors. Vesalius, like Newton a century later, emphasized the phenomena, *i.e.,* the accurate description of natural facts. Vesalius' work touched off a flurry of anatomical work in Italy and elsewhere that culminated in the discovery of the circulation of the blood by William Harvey, whose *Exercitatio Anatomica De Motu Cordis et Sanguinis in Animalibus* (*An Anatomical Exercise Concerning the Motion of the Heart and Blood in Animals*) was published in 1628. This was the *Principia* of physiology that established anatomy and physiology as sciences in their own right. Harvey showed that organic phenomena could be studied experimentally and that some organic processes could be reduced to mechanical systems. The heart and the vascular system could be considered as a pump and a system of pipes and could be understood without recourse to spirits or other forces immune to analysis.

In other sciences the attempt to systematize and criticize was not so successful. In chemistry, for example, the work of the medieval and early modern alchemists had yielded important new substances and processes, such as the mineral acids and distillation, but had obscured theory in almost impenetrable mystical argot. Robert Boyle in England tried to clear away some of the intellectual underbrush by insisting upon clear descriptions, reproducibility of experiments, and mechanical conceptions of chemical processes. Chemistry, however, was not yet ripe for revolution.

In many areas there was little hope of reducing phenomena to comprehensibility, simply because of the sheer number of facts to be accounted for. New instruments like the microscope and the telescope vastly multiplied the worlds with which man had to reckon. The voyages of discovery brought back a flood of new botanical and zoological specimens that overwhelmed ancient classificatory

schemes. The best that could be done was to describe new things accurately and hope that someday they could all be fitted together in a coherent way.

The growing flood of information put heavy strains upon old institutions and practices. It was no longer sufficient to publish scientific results in an expensive book that few could buy; information had to be spread widely and rapidly. Nor could the isolated genius, like Newton, comprehend a world in which new information was being produced faster than any single person could assimilate it. Natural philosophers had to be sure of their data, and to that end they required independent and critical confirmation of their discoveries. New means were created to accomplish these ends. Scientific societies sprang up, beginning in Italy in the early years of the 17th century and culminating in the two great national scientific societies that mark the zenith of the scientific revolution: the Royal Society of London for the Promotion of Natural Knowledge, created by royal charter in 1662, and the Académie des Sciences of Paris, formed in 1666. In these societies and others like them all over the world, natural philosophers could gather to examine, discuss, and criticize new discoveries and old theories. To provide a firm basis for these discussions, societies began to publish scientific papers. The Royal Society's *Philosophical Transactions,* which began as a private venture of its secretary, was the first such professional scientific journal. It was soon copied by the French academy's *Mémoires,* which won equal importance and prestige. The old practice of hiding new discoveries in private jargon, obscure language, or even anagrams gradually gave way to the ideal of universal comprehensibility. New canons of reporting were devised so that experiments and discoveries could be reproduced by others. This required new precision in language and a willingness to share experimental or observational methods. The failure of others to reproduce results cast serious doubts upon the original reports. Thus were created the tools for a massive assault on nature's secrets.

The rise of scientific societies

Even with the scientific revolution accomplished, much remained to be done. Again, it was Newton who showed the way. For the macroscopic world, the *Principia* sufficed. Newton's three laws of motion and the principle of universal gravitation were all that was necessary to analyze the mechanical relations of ordinary bodies, and the calculus provided the essential mathematical tools. For the microscopic world, Newton provided two methods. Where simple laws of action had already been determined from observation, as the relation of volume and pressure of a gas (Boyle's law, $pv = k$), Newton assumed forces between particles that permitted him to derive the law. He then used these forces to predict other phenomena, in this case the speed of sound in air, that could be measured against the prediction. Conformity of observation to prediction was taken as evidence for the essential truth of the theory. Second, Newton's method made possible the discovery of laws of macroscopic action that could be accounted for by microscopic forces. Here the seminal work was not the *Principia* but Newton's masterpiece of experimental physics, the *Opticks,* published in 1704, in which he showed how to examine a subject experimentally and discover the laws concealed therein. Newton showed how judicious use of hypotheses could open the way to further experimental investigation until a coherent theory was achieved. The *Opticks* was to serve as the model in the 18th and early 19th centuries for the investigation of heat, light, electricity, magnetism, and chemical atoms.

THE CLASSIC AGE OF SCIENCE

Mechanics. Just as the *Principia* preceded the *Opticks,* so, too, did mechanics maintain its priority among the sciences in the 18th century, in the process becoming transformed from a branch of physics into a branch of mathematics. Many physical problems were reduced to mathematical ones that proved amenable to solution by increasingly sophisticated analytical methods. The Swiss Leonhard Euler was one of the most fertile and prolific workers in mathematics and mathematical physics. His development of the calculus of variations provided a powerful tool for dealing with highly complex problems. In France, Jean Le Rond d'Alembert and Joseph-Louis Lagrange succeeded in completely mathematizing mechanics, reducing it to an axiomatic system requiring only mathematical manipulation.

The test of Newtonian mechanics was its congruence with physical reality. At the beginning of the 18th century it was put to a rigorous test. Cartesians insisted that the Earth, because it was squeezed at the Equator by the etherial vortex causing gravity, should be somewhat pointed at the poles, a shape rather like that of an American football; Newtonians, arguing that centrifugal force was greatest at the Equator, calculated an oblate sphere that was flattened at the poles and bulged at the Equator. The Newtonians were proved correct after careful measurements of a degree of the meridian were made on expeditions to Lapland and to Peru. The final touch to the Newtonian edifice was provided by Pierre-Simon, marquis de Laplace, whose masterly *Traité de mécanique céleste* (1798–1827; *Celestial Mechanics*) systematized everything that had been done in celestial mechanics under Newton's inspiration. Laplace went beyond Newton by showing that the perturbations of the planetary orbits caused by the interactions of planetary gravitation are in fact periodic and that the solar system is, therefore, stable, requiring no divine intervention.

Chemistry. Although Newton was unable to bring to chemistry the kind of clarification he brought to physics, the *Opticks* did provide a method for the study of chemical phenomena. One of the major advances in chemistry in the 18th century was the discovery of the role of air, and of gases generally, in chemical reactions. This role had been dimly glimpsed in the 17th century, but it was not fully seen until the classic experiments of Joseph Black on *magnesia alba* (basic magnesium carbonate) in the 1750s. By extensive and careful use of the chemical balance, Black showed that an air with specific properties could combine with solid substances like quicklime and could be recovered from them. This discovery served to focus attention on the properties of "air," which was soon found to be a generic, not a specific, name. Chemists discovered a host of specific gases and investigated their various properties: some were flammable, others put out flames; some killed animals, others made them lively. Clearly, gases had a great deal to do with chemistry.

The Newton of chemistry was Antoine-Laurent Lavoisier. In a series of careful balance experiments Lavoisier untangled combustion reactions to show that, in contradiction to established theory, which held that a body gave off the principle of inflammation (called phlogiston) when it burned, combustion actually involves the combination of bodies with a gas that Lavoisier named oxygen. The chemical revolution was as much a revolution in method as in conception. Gravimetric methods made possible precise analysis, and this, Lavoisier insisted, was the central concern of the new chemistry. Only when bodies were analyzed as to their constituent substances was it possible to classify them and their attributes logically and consistently.

Lavoisier

The imponderable fluids. The Newtonian method of inferring laws from close observation of phenomena and then deducing forces from these laws was applied with great success to phenomena in which no ponderable matter figured. Light, heat, electricity, and magnetism were all entities that were not capable of being weighed, *i.e.,* imponderable. In the *Opticks,* Newton had assumed that particles of different sizes could account for the different refrangibility of the various colours of light. Clearly, forces of some sort must be associated with these particles if such phenomena as diffraction and refraction are to be accounted for. During the 18th century heat, electricity, and magnetism were similarly conceived as consisting of particles with which were associated forces of attraction or repulsion. In the 1780s, Charles-Augustin de Coulomb was able to measure electrical and magnetic forces, using a delicate torsion balance of his own invention, and to show that these forces follow the general form of Newtonian universal attraction. Only light and heat failed to disclose such general force laws, thereby resisting reduction to Newtonian mechanics.

Science and the Industrial Revolution. It has long been a commonsensical notion that the rise of modern science

and the Industrial Revolution were closely connected. It is difficult to show any direct effect of scientific discoveries upon the rise of the textile or even the metallurgical industry in Great Britain, the home of the Industrial Revolution, but there certainly was a similarity in attitude to be found in science and nascent industry. Close observation and careful generalization leading to practical utilization were characteristic of both industrialists and experimentalists alike in the 18th century. One point of direct contact is known, namely James Watt's interest in the efficiency of the Newcomen steam engine, an interest that grew from his work as a scientific-instrument maker and that led to his development of the separate condenser that made the steam engine an effective industrial power source. But in general the Industrial Revolution proceeded without much direct scientific help. Yet the potential influence of science was to prove of fundamental importance.

What science offered in the 18th century was the hope that careful observation and experimentation might improve industrial production significantly. In some areas, it did. The potter Josiah Wedgwood built his successful business on the basis of careful study of clays and glazes and by the invention of instruments like the pyrometer with which to gauge and control the processes he employed. It was not, however, until the second half of the 19th century that science was able to provide truly significant help to industry. It was then that the science of metallurgy permitted the tailoring of alloy steels to industrial specifications, that the science of chemistry permitted the creation of new substances, like the aniline dyes, of fundamental industrial importance, and that electricity and magnetism were harnessed in the electric dynamo and motor. Until that period science probably profited more from industry than the other way around. It was the steam engine that posed the problems that led, by way of a search for a theory of steam power, to the creation of thermodynamics. Most importantly, as industry required ever more complicated and intricate machinery, the machine tool industry developed to provide it and, in the process, made possible the construction of ever more delicate and refined instruments for science. As science turned from the everyday world to the worlds of atoms and molecules, electric currents and magnetic fields, microbes and viruses, and nebulae and galaxies, instruments increasingly provided the sole contact with phenomena. A large refracting telescope driven by intricate clockwork to observe nebulae was as much a product of 19th-century heavy industry as were the steam locomotive and the steamship.

The Industrial Revolution had one further important effect on the development of modern science. The prospect of applying science to the problems of industry served to stimulate public support for science. The first great scientific school of the modern world, the École Polytechnique in Paris, was founded in 1794 to put the results of science in the service of France. The founding of scores more technical schools in the 19th and 20th centuries encouraged the widespread diffusion of scientific knowledge and provided further opportunity for scientific advance. Governments, in varying degrees and at different rates, began supporting science even more directly, by making financial grants to scientists, by founding research institutes, and by bestowing honours and official posts on great scientists. By the end of the 19th century the natural philosopher following his private interests had given way to the professional scientist with a public role.

Scientific and technical schools

The Romantic revolt. Perhaps inevitably, the triumph of Newtonian mechanics elicited a reaction, one that had important implications for the further development of science. Its origins are many and complex, and it is possible here to focus on only one, that associated with the German philosopher Immanuel Kant. Kant challenged the Newtonian confidence that the scientist can deal directly with subsensible entities such as atoms, the corpuscles of light, or electricity. Instead, Kant insisted, all that the human mind can know is forces. This epistemological axiom freed Kantians from having to conceive of forces as embodied in specific and immutable particles. It also placed new emphasis on the space between particles; indeed, if one eliminated the particles entirely, there remained only space containing forces. From these two considerations were to come powerful arguments, first, for the transformations and conservation of forces and, second, for field theory as a representation of reality. What makes this point of view Romantic is that the idea of a network of forces in space tied the cosmos into a unity in which all forces were related to all others, so that the universe took on the aspect of a cosmic organism. The whole was greater than the sum of all its parts, and the way to truth was contemplation of the whole, not analysis.

What Romantics, or nature philosophers, as they called themselves, could see that was hidden from their Newtonian colleagues was demonstrated by Hans Christian Ørsted. He found it impossible to believe that there was no connection between the forces of nature. Chemical affinity, electricity, heat, magnetism, and light must, he argued, simply be different manifestations of the basic forces of attraction and repulsion. In 1820 he showed that electricity and magnetism were related, for the passage of an electrical current through a wire affected a nearby magnetic needle. This fundamental discovery was explored and exploited by Michael Faraday, who spent his whole scientific life converting one force into another. By concentrating on the patterns of forces produced by electric currents and magnets, Faraday laid the foundations for field theory, in which the energy of a system was held to be spread throughout the system and not localized in real or hypothetical particles.

The rudiments of field theory

The transformations of force necessarily raised the question of the conservation of force. Is anything lost when electrical energy is turned into magnetic energy, or into heat or light or chemical affinity or mechanical power? Faraday, again, provided one of the early answers in his two laws of electrolysis, based on experimental observations that quite specific amounts of electrical "force" decomposed quite specific amounts of chemical substances. This work was followed by that of James Prescott Joule, Robert Mayer, and Hermann von Helmholtz, each of whom arrived at a generalization of basic importance to all science, the principle of the conservation of energy.

The nature philosophers were primarily experimentalists who produced their transformations of forces by clever experimental manipulation. The exploration of the nature of elemental forces benefitted as well from the rapid development of mathematics. In the 19th century the study of heat was transformed into the science of thermodynamics, based firmly on mathematical analysis; the Newtonian corpuscular theory of light was replaced by Augustin-Jean Fresnel's mathematically sophisticated undulatory theory; and the phenomena of electricity and magnetism were distilled into succinct mathematical form by William Thomson (Lord Kelvin) and James Clerk Maxwell. By the end of the century, thanks to the principle of the conservation of energy and the second law of thermodynamics, the physical world appeared to be completely comprehensible in terms of complex but precise mathematical forms describing various mechanical transformations in some underlying ether.

The submicroscopic world of material atoms became similarly comprehensible in the 19th century. Beginning with John Dalton's fundamental assumption that atomic species differ from one another solely in their weights, chemists were able to identify an increasing number of elements and to establish the laws describing their interactions. Order was established by arranging elements according to their atomic weights and their reactions. The result was the periodic table, devised by Dmitry Mendeleyev, which implied that some kind of subatomic structure underlay elemental qualities. That structure could give rise to qualities, thus fulfilling the prophecy of the 17th-century mechanical philosophers, was shown in the 1870s by Joseph-Achille Le Bel and Jacobus van't Hoff, whose studies of organic chemicals showed the correlation between the arrangement of atoms or groups of atoms in space and specific chemical and physical properties.

The founding of modern biology. The study of living matter lagged far behind physics and chemistry, largely because organisms are so much more complex than inanimate bodies or forces. Harvey had shown that living

matter could be studied experimentally, but his achievement stood alone for two centuries. For the time being, most students of living nature had to be content to classify living forms as best they could and to attempt to isolate and study aspects of living systems.

As has been seen, an avalanche of new specimens in both botany and zoology put severe pressure on taxonomy. A giant step forward was taken in the 18th century by the Swedish naturalist Carl von Linné—known by his Latinized name, Linnaeus—who introduced a rational, if somewhat artificial, system of binomial nomenclature. The very artificiality of Linnaeus' system, focussing as it did on only a few key structures, encouraged criticism and attempts at more natural systems. The attention thus called to the organism as a whole reinforced a growing intuition that species are linked in some kind of genetic relationship, an idea first made scientifically explicit by Jean-Baptiste, chevalier de Lamarck.

Problems encountered in cataloging the vast collection of invertebrates at the Museum of Natural History in Paris led Lamarck to suggest that species change through time. This idea was not so revolutionary as it is usually painted, for, although it did upset some Christians who read the book of Genesis literally, naturalists who noted the shading of natural forms one into another had been toying with the notion for some time. Lamarck's system failed to gain general assent largely because it relied upon an antiquated chemistry for its causal agents and appeared to imply a conscious drive to perfection on the part of organisms. It was also opposed by one of the most powerful paleontologists and comparative anatomists of the day, Georges Cuvier, who happened to take Genesis quite literally. In spite of Cuvier's opposition, however, the idea remained alive and was finally elevated to scientific status by the labours of Charles Darwin. Darwin not only amassed a wealth of data supporting the notion of transformation of species, but he also was able to suggest a mechanism by which such evolution could occur without recourse to other than purely natural causes. The mechanism was natural selection, according to which minute variations in offspring were either favoured or eliminated in the competition for survival, and it permitted the idea of evolution to be perceived with great clarity. Nature shuffled and sorted its own productions, through processes governed purely by chance, so that those organisms that survived were better adapted to a constantly changing environment.

Darwin

Darwin's *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life,* published in 1859, brought order to the world of organisms. A similar unification at the microscopic level had been brought about by the cell theory announced by Theodor Schwann and Matthias Schleiden in 1838, whereby cells were held to be the basic units of all living tissues. Improvements in the microscope during the 19th century made it possible gradually to lay bare the basic structures of cells, and rapid progress in biochemistry permitted the intimate probing of cellular physiology. By the end of the century the general feeling was that physics and chemistry sufficed to describe all vital functions and that living matter, subject to the same laws as inanimate matter, would soon yield up its secrets. This reductionist view was triumphantly illustrated in the work of Jacques Loeb, who showed that so-called instincts in lower animals are nothing more than physicochemical reactions, which he labelled tropisms.

The most dramatic revolution in 19th-century biology was the one created by the germ theory of disease, championed by Louis Pasteur in France and Robert Koch in Germany. Through their investigations, bacteria were shown to be the specific causes of many diseases. By means of immunological methods first devised by Pasteur, some of mankind's chief maladies were brought under control.

THE 20TH-CENTURY REVOLUTION

By the end of the 19th century, the dream of the mastery of nature for the benefit of mankind, first expressed in all its richness by Sir Francis Bacon, seemed on the verge of realization. Science was moving ahead on all fronts, reducing ignorance and producing new tools for the amelioration of the human condition. A comprehensible, rational view of the world was gradually emerging from laboratories and universities. One savant went so far as to express pity for those who would follow him and his colleagues, for they, he thought, would have nothing more to do than to measure things to the next decimal place.

But this sunny confidence did not last long. One annoying problem was that the radiation emitted by atoms proved increasingly difficult to reduce to known mechanical principles. More importantly, physics found itself relying more and more upon the hypothetical properties of a substance, the ether, that stubbornly eluded detection. Within a span of 10 short years, roughly 1895–1905, these and related problems came to a head and wrecked the mechanistic system the 19th century had so laboriously built. The discovery of X rays and radioactivity revealed an unexpected new complexity in the structure of atoms. Max Planck's solution to the problem of thermal radiation introduced a discontinuity into the concept of energy that was inexplicable in terms of classical thermodynamics. Most disturbing of all, the enunciation of the special theory of relativity by Albert Einstein in 1905 not only destroyed the ether and all the physics that depended on it but also redefined physics as the study of relations between observers and events, rather than of the events themselves. What was observed, and therefore what happened, was now said to be a function of the observer's location and motion relative to other events. Absolute space was a fiction. The very foundations of physics threatened to crumble.

Einstein

This modern revolution in physics has not yet been fully assimilated by historians of science. Suffice it to say that scientists managed to come to terms with all of the upsetting results of early 20th-century physics but in ways that made the new physics utterly different from the old. Mechanical models were no longer acceptable, because there were processes (like light) for which no consistent model could be constructed. No longer could physicists speak with confidence of physical reality, but only of the probability of making certain measurements.

All this being said, there is still no doubt that science in the 20th century has worked wonders. The new physics—relativity, quantum mechanics, particle physics—may outrage common sense, but it enables physicists to probe to the very limits of physical reality. Their instruments and mathematics permit modern scientists to manipulate subatomic particles with relative ease, to reconstruct the first moment of creation, and to glimpse dimly the grand structure and ultimate fate of the universe.

The revolution in physics has spilled over into chemistry and biology and led to hitherto undreamed-of capabilities for the manipulation of atoms and molecules and of cells and their genetic structures. Chemists perform molecular tailoring today as a matter of course, cutting and shaping molecules at will. Genetic engineering makes possible active human intervention in the evolutionary process and holds out the possibility of tailoring living organisms, including the human organism, to specific tasks. This second scientific revolution may prove to be, for good or ill, the most important event in the history of mankind.

(L.P.W.)

BIBLIOGRAPHY

General works. GEORGE SARTON, *A History of Science,* 2 vol. (1952–59, reissued 1993), and *Introduction to the History of Science,* 3 vol. in 5 (1927–48, reprinted 1975), embody the legacy of this founder of the discipline of the history of science. Though more philosophical in tone, the immensely influential work by THOMAS KUHN, *The Structure of Scientific Revolutions,* 3rd ed. (1996), is essential reading. Of several surveys of the entire field, J.D. BERNAL, *Science in History,* new ed., 4 vol. (1969, reissued 1979), although Marxist-inflected, is the most easily accessible and useful—vol. 1 and 2 are the strongest. DAVID C. LINDBERG, *The Beginnings of Western Science* (1992), is the best recent overview of Western science to 1400, with a thorough bibliography. JOSEPH NEEDHAM *et al., Science and Civilisation in China* (1954–), is the indispensable history of Chinese science and technology; COLIN RONAN, *The Shorter Science and Civilisation in China* (1980–), is an abridgment.

General studies of specific fields include DIRK J. STRUIK, *A Concise History of Mathematics,* 4th rev. ed. (1987); on astronomy, TIMOTHY FERRIS, *Coming of Age in the Milky Way* (1988); and JOHN NORTH, *The Fontana History of Astronomy and Cos-*

mology (also published as *The Norton History of Astronomy and Cosmology,* 1994); PETER J. BOWLER, *The Fontana History of the Environmental Sciences* (1992; also published as *The Norton History of the Environmental Sciences,* 1993), a history of biological and ecological thought; and WILLIAM H. BROCK, *The Fontana History of Chemistry* (1992; also published as *The Norton History of Chemistry,* 1993). The relationship between science and religion is surveyed in JOHN HEDLEY BROOKE, *Science and Religion: Some Historical Perspectives* (1991).

CHARLES COULTON GILLISPIE (ed.), *Dictionary of Scientific Biography,* 16 vol. (1970–80), is the definitive biographical reference source for the field. ROY PORTER (ed.), *The Biographical Dictionary of Scientists,* 2nd ed. (1994), offers briefer biographies. *Isis Current Bibliography of the History of Science and Its Cultural Influences* (annual) surveys the most recent literature.

Ongoing research is reported in a number of journals. *Isis* (quarterly) is the leading U.S. journal; and *Osiris* (annual) is also published in the United States; while *History of Science* (quarterly) and *The British Journal for the History of Science* (quarterly) are good British journals with wide coverage, bibliographies, and essay reviews.

Ancient and medieval science. MOTT T. GREENE, *Natural Knowledge in Preclassical Antiquity* (1992), is a general survey of preclassical sciences; while O. NEUGEBAUER, *The Exact Sciences in Antiquity,* 2nd ed. (1957, reissued 1993), focusses on ancient mathematics and astronomy. The best general history of Greek science is G.E.R. LLOYD, *Early Greek Science: Thales to Aristotle* (1970), and *Greek Science After Aristotle* (1973). Roman science is treated in WILLIAM H. STAHL, *Roman Science: Origins, Development, and Influence to the Later Middle Ages* (1962, reprinted 1978).

There is no book-length survey of Islāmic science, but A.I. SABRA, "Science, Islamic," in JOSEPH R. STRAYER (ed.), *Dictionary of the Middle Ages* (1988), vol. 11, pp. 81–89, is a good short overview. Astronomy and physics are discussed in A.I. SABRA, *Optics, Astronomy, and Logic: Studies in Arabic Science and Philosophy* (1994); DAVID A. KING, *Astronomy in the Service of Islam* (1993); and EDWARD GRANT, *Planets, Stars, and Orbs: The Medieval Cosmos, 1200–1687* (1994). TOBY E. HUFF, *The Rise of Early Modern Science* (1993), compares science in the medieval Islāmic world, China, and the West. A.C. CROMBIE, *Augustine to Galileo,* 2nd rev. ed., 2 vol. (1959, reissued as *The History of Science from Augustine to Galileo,* 2 vol. in 1, 1995), is still a useful introduction to medieval science. EDWARD GRANT, *Physical Science in the Middle Ages* (1971); and MARSHALL CLAGETT, *The Science of Mechanics in the Middle Ages* (1959, reissued 1979), discuss medieval physics. LYNN WHITE, "The Historical Roots of Our Ecological Crisis," *Science,* 155:1203–07 (March 10, 1967), presents an influential argument about medieval attitudes toward nature.

The scientific revolution. ALLEN G. DEBUS, *Man and Nature in the Renaissance* (1978); and STEVEN SHAPIN, *The Scientific Revolution* (1996), are the best short overviews. Additional surveys include MARGARET C. JACOB, *The Cultural Meaning of the Scientific Revolution* (1988, reissued 1993), strong on comparative and social issues; and DAVID C. LINDBERG and ROBERT S. WESTMAN (eds.), *Reappraisals of the Scientific Revolution* (1990), a cross section of current scholarship. STEVEN SHAPIN and SIMON SCHAFFER, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (1985); and STEVEN SHAPIN, *A Social History of Truth: Civility and Science in Seventeenth-Century England* (1994), trace the development of the experimental method. SAMUEL Y. EDGERTON, JR., *The Heritage of Giotto's Geometry* (1991), explores the relationship between art and science in the Renaissance. Physical science and astronomy are discussed in ALEXANDRE KOYRÉ, *From the Closed World to the Infinite Universe* (1957, reissued 1994); THOMAS S. KUHN, *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (1957, reissued 1985); and RICHARD S. WESTFALL, *The Construction of Modern Science: Mechanisms and Mechanics* (1971). Natural history and biology are treated in MARTIN J.S. RUDWICK, *The Meaning of Fossils: Episodes in the History of Palaeontology,* 2nd rev. ed. (1976, reprinted 1985). The links between Hermeticism, magic, and science are explored in FRANCES A. YATES, *Giordano Bruno and the Hermetic Tradition* (1964, reissued 1991). Alchemy is examined in BETTY JO TEETER DOBBS, *The Janus Faces of Genius: The Role of Alchemy in Newton's Thought* (1991); WILLIAM R. NEWMAN, *Gehennical Fire: The Lives of George Starkey, an American Alchemist in the Scientific Revolution* (1994); and PAMELA H. SMITH, *The Business of Alchemy: Science and Culture in the Holy Roman Empire* (1994). Studies of gender and Renaissance science include EVELYN FOX KELLER, *Reflections on Gender and Science* (1985, reissued 1995), an essential starting point; and the controversial work by DAVID F. NOBLE, *A World Without Women: The Christian Clerical Culture of Western Science* (1992).

Modern science. THOMAS L. HANKINS, *Science and the Enlightenment* (1985), is an excellent survey of 18th-century science. JOSEPH BEN-DAVID, *The Scientist's Role in Society* (1971, reprinted with a new introduction, 1984), surveys the development of scientific institutions and communities in Europe and America. MARGARET C. JACOB (ed.), *The Politics of Western Science, 1640–1990* (1994); and PETER GALISON and BRUCE HEVLY (eds.), *Big Science: The Growth of Large-Scale Research* (1992), are useful introductions to the politics of science. Works on gender and science include MARGARET ALIC, *Hypatia's Heritage: A History of Women in Science from Antiquity Through the Nineteenth Century* (1986), which is strongest on the modern period; LONDA L. SCHIEBINGER, *Nature's Body: Gender in the Making of Modern Science* (1993); LUDMILLA JORDANOVA, *Sexual Visions: Images of Gender in Science and Medicine Between the Eighteenth and Twentieth Centuries* (1989, reissued 1993); and MARGARET W. ROSSITER, *Women Scientists in America: Struggles and Strategies to 1940* (1982), and *Women Scientists in America: Before Affirmative Action, 1940–1972* (1995). Surveys of science in the national and imperial context include NANCY STEPAN, *Beginnings of Brazilian Science* (1976, reissued 1981); LEWIS PYENSON, *Cultural Imperialism and the Exact Sciences: German Expansion Overseas, 1900–1930* (1985); DANIEL R. HEADRICK, *The Tentacles of Progress: Technology Transfer in the Age of Imperialism, 1850–1940* (1988); JAMES R. BARTHOLOMEW, *The Formation of Science in Japan* (1989); and LOREN R. GRAHAM, *Science in Russia and the Soviet Union* (1993).

The literature on specific sciences since the 18th century is voluminous. The physical sciences are discussed in J.L. HEILBRON, *Electricity in the 17th and 18th Centuries* (1979); CHRISTA JUNGNICKEL and RUSSELL MCCORMMACH, *Intellectual Mastery of Nature: Theoretical Physics from Ohm to Einstein,* 2 vol. (1986), strong on the late 19th century; DANIEL J. KEVLES, *The Physicists* (1978, reprinted 1995), a history of American physics; P.M. HARMAN, *Energy, Force, and Matter* (1982), on 19th-century developments; and three important collections of essays: PETER GALISON, *How Experiments End* (1987), on high-energy physics; DAVID GOODING, TREVOR PINCH, and SIMON SCHAFFER (eds.), *The Uses of Experiment: Studies in the Natural Sciences* (1989); and JED Z. BUCHWALD (ed.), *Scientific Practice: Theories and Stories of Doing Physics* (1995). HENRY C. KING, *The History of the Telescope* (1955, reprinted 1979), is still useful as a history of astronomy; while OWEN GINGERICH (ed.), *Astrophysics and Twentieth-Century Astronomy to 1950,* vol. 1 (1984), is a more recent survey. Quantification and mathematics are discussed in TORE FRÄNGSMYR, J.L. HEILBRON, and ROBIN E. RIDER (eds.), *The Quantifying Spirit in the 18th Century* (1990); LORRAINE DASTON, *Classical Probability in the Enlightenment* (1988); THEODORE M. PORTER, *The Rise of Statistical Thinking, 1820–1900* (1986), and *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (1995); and JOAN L. RICHARDS, *Mathematical Visions: The Pursuit of Geometry in Victorian England* (1988). Voyages of discovery are studied in HARRY WOOLF, *The Transits of Venus: A Study of Eighteenth-Century Science* (1959); BERNARD SMITH, *European Vision and the South Pacific,* 2nd ed. (1985), and *Imagining the Pacific* (1992); and WILLIAM H. GOETZMANN, *New Lands, New Men: America and the Second Great Age of Discovery* (1986, reissued 1995). Life sciences are discussed in DAVID ELLISTON ALLEN, *The Naturalist in Britain* (1976, reissued 1994), a social history of natural history and collecting; GARLAND E. ALLEN, *Life Science in the Twentieth Century* (1975), a general survey; and RONALD RAINGER, KEITH R. BENSON, and JANE MAIENSCHEIN (eds.), *The American Development of Biology* (1988, reissued 1991). Evolution is the subject of PETER J. BOWLER, *Evolution: The History of an Idea,* rev. ed. (1989); and ADRIAN DESMOND, *The Politics of Evolution: Morphology, Medicine, and Reform in Radical London* (1989, reissued 1992). Genetics and molecular biology are discussed in HORACE FREELAND JUDSON, *The Eighth Day of Creation: Makers of the Revolution in Biology* (1979); JAMES D. WATSON, *The Double Helix* (1968), available also in a critical edition ed. by GUNTHER S. STENT (1980), an interesting and controversial autobiographical account of the discovery of DNA structure; ADELE E. CLARKE and JOAN H. FUJIMURA (eds.), *The Right Tools for the Job* (1992), on instruments and animals in 20th-century biology; and ROBERT E. KOHLER, *From Medical Chemistry to Biochemistry* (1982), and *Lords of the Fly* (1994), on *Drosophila melanogaster* and genetics research. The social sciences are discussed in CHRISTOPHER FOX, ROY PORTER, and ROBERT WOKLER (eds.), *Inventing Human Science: Eighteenth-Century Domains* (1995); DOROTHY ROSS, *The Origins of American Social Science* (1991); DANIEL J. KEVLES, *In the Name of Eugenics: Genetics and the Uses of Human Heredity* (1985, reissued 1995); GEORGE W. STOCKING, JR., *Victorian Anthropology* (1987); and DONNA HARAWAY, *Primate Visions: Gender, Race, and Nature in the World of Modern Science* (1989), a history of primatology.

(L.P.W./Ed.)

Science Fiction

Science fiction, often abbreviated SF or sci-fi, deals principally with the impact of actual or imagined science upon society or individuals. The term *science fiction* was popularized in the 1920s by one of the genre's principal advocates, the American publisher Hugo Gernsback. The Hugo Awards, given annually since 1953 by the World Science Fiction Society, recognize the top SF writers, editors, illustrators, films, and "fanzines."

Though writers in antiquity sometimes dealt with themes common to modern science fiction, their stories made no attempt at scientific and technological plausibility, the feature that distinguishes science fiction from earlier speculative writings and other contemporary speculative genres such as fantasy and horror. The genre formally emerged in the West, where the social transformations wrought by the Industrial Revolution first led writers and intellectuals to extrapolate the future impact of technology. By the beginning of the 20th century, an array of standard SF "sets" had developed around certain themes, among them space travel, robots, alien beings, and time travel. The customary "theatrics" of science fiction include prophetic warnings, utopian aspirations, elaborate scenarios for entirely imaginary worlds, titanic disasters, strange voyages, and political agitation of many extremist flavours, presented in the form of sermons, meditations, satires, allegories, and parodies—exhibiting every conceivable attitude toward the process of techno-social change, from cynical despair to cosmic bliss.

SF writers often seek out new scientific and technical developments in order to prognosticate freely the techno-social changes that will shock the readers' sense of cultural propriety and expand their consciousness. This approach was central to the work of H.G. Wells, a founder of the genre and likely its greatest writer. Wells was an ardent student of the 19th-century British scientist T.H. Huxley, whose vociferous championing of Charles Darwin's theory of evolution earned him the epithet "Darwin's Bulldog." Wells's work gives ample evidence of SF's latent radicalism and its affinity for aggressive satire and utopian political agendas, as well as its dire predictions of technological destruction.

This dark side can be seen in the work of T.H. Huxley's grandson, Aldous Huxley, who was a social satirist, an advocate of psychedelic drugs, and the author of a dystopian classic, *Brave New World* (1932). The sense of dread was also cultivated by H.P. Lovecraft, who invented the *Necronomicon*, an imaginary book of knowledge so ferocious that any scientist who dares to read it succumbs to madness. On a more personal level, the works of Philip K. Dick present metaphysical conundrums about identity, humanity, and the nature of reality. Perhaps bleakest of all, the English philosopher Olaf Stapledon's mind-stretching novels picture all of human history as a frail, passing bubble in the cold galactic stream of space and time.

When the genre began to gel in the early 20th century, it was generally disreputable, particularly in the United States, where it first catered to a juvenile audience. Following World War II, science fiction spread throughout the world from its epicentre in the United States, spurred on by ever more staggering scientific feats—*e.g.*, the development of the atomic bomb, human visits to the Moon, and the real possibility of cloning human life.

By the 21st century, science fiction had become much more than a literary genre. Its avid followers and practitioners constitute a thriving worldwide subculture. Fans relish the seemingly endless variety of SF-related products and frequently hold well-attended, well-organized conventions, at which costumes are worn.

This article provides a broad survey of the historical development of science fiction. Notable examples of science fiction in film and television are also discussed in the article, which is divided into the following sections:

The evolution of science fiction 42A
- Antecedents 42A
- The 19th and early 20th centuries 42B
  - Proto-science fiction
  - Jules Verne
  - Classic British science fiction
  - Mass markets and juvenile science fiction
  - The "golden age" of science fiction
  - Soviet science fiction
- Science fiction after World War II 42C
  - New directions in fiction
  - SF cinema and TV

Major science fiction themes 42D
- Utopias and dystopias 42D
- Alternative societies 42E
- Sex and gender 42E
- Alien encounters 42F
- Space travel 42F
- Time travel 42G
- Alternate histories and parallel universes 42G
- High technologies 42H

Bibliography 42H

The evolution of science fiction

ANTECEDENTS

Antecedents of science fiction can be found in the remote past. Among the earliest examples is the 2nd-century-AD Syrian-born Greek satirist Lucian, who in *Trips to the Moon* describes sailing to the Moon. Such flights of fancy, or fantastic tales, provided a popular format in which to satirize government, society, and religion while evading libel suits, censorship, and persecution. The clearest forerunner of the genre, however, was the 17th-century swashbuckler Cyrano de Bergerac, who wrote of a voyager to the Moon finding a utopian society of men free from war, disease, and hunger. (See below *Utopias and dystopias*.) The voyager eats fruit from the biblical tree of knowledge and joins lunar society as a philosopher—that is, until he is expelled from the Moon for blasphemy. Following a short return to Earth, he travels to the Sun, where a society of birds puts him on trial for humanity's crimes. In creating his diversion, Cyrano took it as his mission to make impossible things seem plausible. Although this and his other SF-like writings were published only posthumously and in various censored versions, Cyrano had a great influence on later satirists and social critics. Two works in particular—Jonathan Swift's *Gulliver's Travels* (1726) and Voltaire's *Micromégas* (1752)—show Cyrano's mark with their weird monsters, gross inversions of normalcy, and harsh satire.

Another precursor was Louis-Sébastien Mercier's *L'An deux mille quatre cent quarante* (*c.* 1771; "The Year 2440"; *Memoirs of the Year Two Thousand Five Hundred*), a work of French political speculation set in a 25th-century utopian society that worships science. While many writers had depicted some future utopian "Kingdom of God" or a utopian society in some mythical land, this was the first work to postulate a utopian society on Earth in the realizable future. The book was swiftly banned by the French ancien régime, which recognized that Mercier's fantasy about "the future" was a thin disguise for his subversive revolutionary sentiments. Despite this official sanction—or perhaps because of it—Mercier's book became an international best-seller. Both Thomas Jefferson and George Washington owned copies.

THE 19TH AND EARLY 20TH CENTURIES

Proto-science fiction. In 1818 Mary Wollstonecraft Shelley took the next major step in the evolution of science fiction when she published *Frankenstein; or, The Modern Prometheus*. Champions of Shelley as the "mother of science fiction" emphasize her innovative fictional scheme. Abandoning the occult folderol of the conventional Gothic novel, she made her protagonist a practicing "scientist"—though the term *scientist* was not actually coined until 1834—and gave him an interest in galvanic electricity and vivisection, two of the advanced technologies of the early 1800s. Even though reanimated corpses remain fantastic today, Shelley gave her story an air of scientific plausibility. This masterly manipulation of her readers established a powerful new approach to creating thrilling sensations of wonder and fear. *Frankenstein* has been adapted for film repeatedly since the first silent version in 1910. Frankenstein's monster likewise remained a potent metaphor at the turn of the 21st century, when opponents of genetically engineered food coined the term *Frankenfood* to express their concern over the unknown effects of the human manipulation of foodstuffs.

Frankenstein

Another significant 19th-century forerunner was Edgar Allan Poe, who wrote many works loosely classifiable as science fiction. "The Balloon Hoax" of 1844, originally published in the *New York Sun*, is but one example of Poe's ability to provide meticulous technical descriptions intended to mislead and impress the gullible.

Jules Verne. More significant to the genre's formation than Poe was Jules Verne, who counted Poe among his influences and was arguably the inventor of science fiction. Verne's first novel, *Paris au XXième siècle* (*Paris in the Twentieth Century*)—written in 1863 but not published until 1994—is set in the distant 1960s and contains some of his most accurate prognostications: elevated trains, automobiles, facsimile machines, and computer-like banking machines. Nevertheless, the book's depiction of a dark and bitter dystopian world without art was too radical for Jules Hetzel, Verne's publisher.

Hetzel, who published a popular-science magazine for young people, the *Magasin illustré d'éducation et de récréation*, was a shrewder judge of public taste than Verne. With Hetzel's editorial guidance, Verne abandoned his farfetched futurism and set to work on the first of his *Voyages extraordinaires—Cinq Semaines en ballon* (1863; *Five Weeks in a Balloon*). In this series of contemporary techno-thrillers, the reader learns of balloons, submarines, trains, mechanical elephants, and many other engineering marvels, all described with unmatchable technical accuracy and droll humour.

Verne's novels achieved remarkable international success, and he became a legend in his own time. His major works—often adapted for film—remained popular into the 21st century, and the "scientific romance" became a permanent fixture of Western popular entertainment.

Another uncannily prescient figure was the French illustrator Albert Robida. His graphic cartoons and essays appeared in *Le Vingtième Siècle* (1882; "The 20th Century"), *La Vie électrique* (1883; "The Electric Life"), and the particularly ominous and impressive *La Guerre au vingtième siècle* (1887; "War in the 20th Century"). Although Robida's shrewd extrapolations were created for comic effect, they proved remarkably akin to the 20th century's reality. In fact, since Robida's time, science fiction has often proved most prophetic not at its magisterial heights of moral sobriety but at its most louche and peculiar.

Classic British science fiction. Great Britain as well as France experienced a flowering of creative imagination in the 1880s and '90s. Literary landmarks of the period included such innovative works as Robert Louis Stevenson's *Strange Case of Dr. Jekyll and Mr. Hyde* (1886) and H.G. Wells's phenomenal trio of *The Time Machine* (1895), *The Invisible Man* (1897), and *The War of the Worlds* (1898). As the 20th century dawned, many of science fiction's most common themes—space travel, time travel, utopias and dystopias, and encounters with alien beings—bore British postmarks.

The technophilic tenor of the times, as well as 19th-century laissez-faire capitalism, also inspired a reaction from those who longed for a return to a preindustrial life. William Morris' *News from Nowhere* (1890) envisioned a 21st-century pastoral utopia that combined the author's socialist theories with the lucid and placid values of the 14th century. While some critics dismissed Morris' work as a communist tract, C.S. Lewis praised its style and language. Indeed, Lewis, Lord Dunsany, E.R. Eddison, J.R.R. Tolkien, and a growing host of imitators imbued pastoral settings with heroic and mythic elements, often borrowing from Christian ethos. Examples of this type of work existed even across the Atlantic, notably in two novels by William Dean Howells, the dean of late 19th-century American letters. In Howells' *A Traveler from Altruria* (1894) and *Through the Eye of the Needle* (1907), he described Altruria, a utopian world that combined the foundations of Christianity and the U.S. Constitution to produce an "ethical socialism" by which society was guided. Though heroic fantasy remained a minority taste for many decades, during the second half of the 20th century it began to dominate bookstore shelves and book clubs (see *Science fiction after World War II*).

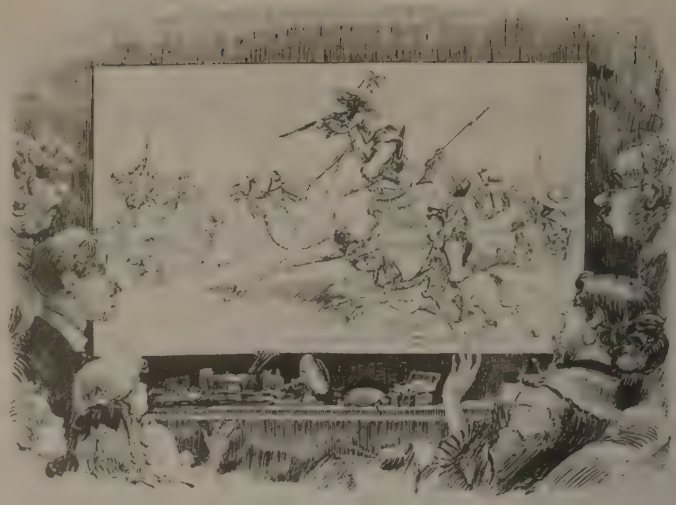


Illustration from the 1880s, in which Albert Robida imagined news coverage in the 20th century, anticipating flat-screen, home-theatre television and live video news broadcasts.
Photograph, Christine E. Haycock, M.D.

Altruria

Mass markets and juvenile science fiction. Publishing trends brought about an important shift in the development of the genre. The most crucial change in Britain was a decline in the publication of "three-decker" Victorian novels and an accompanying expansion of magazine publication. This adjustment proved highly advantageous to shorter works of science fiction. It brought about a new subgenre, as seen, for example, in George Chesney's short story "The Battle of Dorking" (1871), which darkly postulated a Prussian defeat of a poorly armed, weak, and unwary Britain and established the military techno-thriller. Chesney used his urgent narrative of the near future to warn against Britain's decline.

Magazine publication was encouraged by an even more pronounced publishing trend that began in the early 1880s. With the development of a cheap process for converting wood pulp into paper and the increasing mechanization of the printing process, inexpensive "pulp" magazines began to deliver stories to a mass audience. During this period in the United States, "dime novels" (shoddily produced pamphlets that usually sold for a nickel) and boys' adventure magazines proliferated. The stories distributed in these books and magazines, such as Luis Senarens' *Frank Reade, Jr., and His Steam Wonder* (1884), often boasted SF elements that appealed to the young reader's sense of wonder and adventure. While Verne's influence is evident in them, dime novels lacked both Verne's knowledge of technology and his literary skill. Senarens' work, for example, epitomizes the worst aspects of the type: they are poorly written and filled with sadistic racism directed toward Native Americans, African Americans, Irish Americans, Mexicans, and Jews.

Pulp fiction

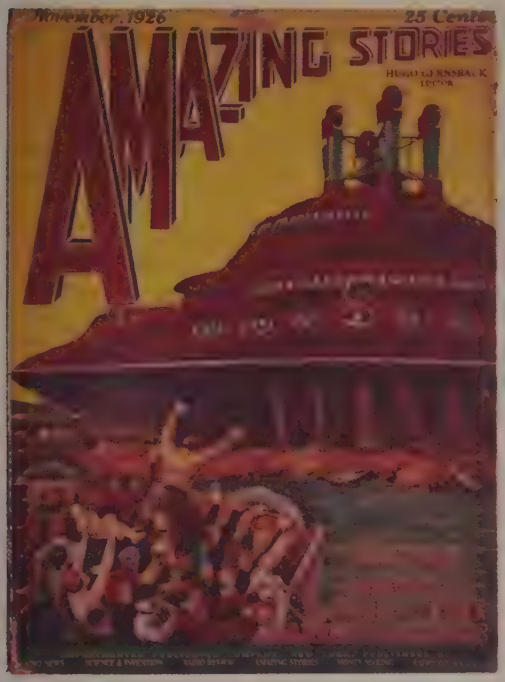
Edgar Rice Burroughs, with his serialized story "Under the Moons of Mars" (1912; novelized as *A Princess of Mars* and *Under the Moons of Mars*), transformed European-style "literary" science fiction into a distinctly American genre directed at a juvenile audience. Combining European elements of fantasy and horror with the naive expansionist style of early American westerns, Burroughs had his hero John Carter outwit various inferior green, yellow, and black Martians. He also marries a red Martian and has a child by her, despite the fact that she reproduces by laying eggs. Burroughs's hero remained a SF archetype, especially for "space operas," through the 1950s.

The success of such stories inculcated a love of science fiction that culminated in the founding of adult-oriented SF pulp magazines in the 1920s.

The "golden age" of science fiction. The previously mentioned Hugo Gernsback, an emigrant from Luxembourg based in New York City, made a living publishing technical magazines for radio and electrical enthusiasts. Noting the growing fondness of his youthful audience for fictional accounts of thrilling technical wonders, Gernsback began to republish the works of Verne and Poe and the early writings of H.G. Wells in great profusion.

Amazing Stories

Gernsback's magazine *Amazing Stories* (founded 1926) broke ground for many imitators and successors, including his own later periodicals *Science Wonder Stories*, *Air Wonder Stories*, and *Scientific Detective Monthly* (later known as *Amazing Detective Tales*), and a torrent of other pulp publications. This practice soon yielded so much fruit that many people, especially Americans, falsely assumed that Americans had created science fiction.



The November 1926 issue of *Amazing Stories*, which featured a new story by Garrett P. Serviss, with cover art and illustrations by Frank R. Paul. It also reprinted works of H.G. Wells and Jules Verne.
Photograph, Christine E. Haycock, M.D

By 1934 SF readership in the United States was large enough to support the establishment of the Science Fiction League, Gernsback's professionally sponsored fan organization (with local chapters in the United Kingdom and Australia). Like a kind of freemasonry, SF fandom spread across the United States. Eager young devotees soon had their own stories published, and, as time passed, they became the hardened, canny professionals of the SF pulp world. Literary groups such as New York's Futurians, Milwaukee's Fictioneers, and the Los Angeles Science Fiction League argued ideology in amateur presses. Conventions were held, feuds and friendships flourished, and science fiction began its long climb, never to respectability but rather toward mass acceptance.

Another influential figure was John W. Campbell, Jr., who from 1937 to 1971 edited *Astounding Science Fiction*. Campbell's insistence on accurate scientific research (he attended the Massachusetts Institute of Technology and received his B.S. in physics from Duke University) and some sense of literary style shaped the career of almost every major American science fiction writer from the period. As a writer, Campbell is noteworthy for his story "Who Goes There?" (1938) and its two film versions, *The Thing* (1951 and 1982), but he is best remembered as an editor. Many fans refer to Campbell's early years at *Astounding*, roughly 1938–46, with its frequent publication of stories by Robert Heinlein, Isaac Asimov, A.E. Van Vogt, and Theodore Sturgeon, as SF's golden age.

Certain literary critics countered wittily that the "golden age" of science fiction is the chronological age of 14—the reputed age at which many fans become hooked on science fiction and the all-too-typical literary level of a genre relished far more for its new scientific "ideas" than its literary merits. Nevertheless, even the sharpest critic would have to admit that for all its often juvenile nature—particularly as conceived in the United States—science fiction was a singular source of scientific wonder and discovery that inspired generations of scientists and engineers to pursue in reality what they had dreamed about in their youth.

Soviet science fiction. Only the gargantuan world of Soviet state publishing could match the production of U.S. science fiction. The Soviet promotion of "scientific socialism" created a vital breathing space for science fiction within Soviet society. The genre's often allegorical nature gave Soviet writers of science fiction many creative opportunities for relatively free expression.

Soviet science fiction was broad and deep enough to spawn several subgenres, such as the techno-thriller Red Detective stories of Marxist world revolution and many cosmonaut space operas. Among its masterpieces were the Constructivist silent film *Aelita* (1924), based on the 1923 novel of the same title by Aleksey Tolstoy. The film's imaginative set and costume designs had a strong artistic influence on Fritz Lang's film *Metropolis* (1927). Both *Aelita*'s design and its scenes of an Earthman leading a Martian proletarian revolt against an oppressive regime were echoed in the 1930s American film serial *Flash Gordon*. Another notable work of this period was Yevgeny Zamyatin's *My* (written in 1920, circulated in manuscript and not published in Russian until 1952; translated into English as *We* in 1924), which won a wide readership overseas, though the author's satiric daring led to his banishment under Joseph Stalin. The book's depiction of life under a totalitarian state influenced the other two great dystopian novels of the 20th century, Aldous Huxley's *Brave New World* (1932) and George Orwell's *Nineteen Eighty-four* (1949).

Aelita

SCIENCE FICTION AFTER WORLD WAR II

New directions in fiction. After World War II, publishers largely abandoned the pulps in favour of paperback books and paperback-like "digests." By that time, however, science fiction had inspired such passionate devotion that it moved with ease into small specialty presses. Two new digest magazines in particular—*The Magazine of Fantasy and Science Fiction* (1949–) and *Galaxy Science Fiction* (1950–80)—prospered. Science fiction also grew in popular esteem after the advent of the atomic bomb (1945) and the launch of Sputnik (1957).

Under the editorial guidance of these new digests, American SF of the 1950s became more sophisticated, urbane, and satiric, with raw technophilia waning in favour of more anthropologically based speculation about societies and cultures. Many books (and film adaptations) from the decade were rife with Cold War-induced fear and paranoia. Perhaps the most representative novel is Walter M. Miller's *A Canticle for Leibowitz* (1960; first serialized 1955–57), which describes the post-nuclear-holocaust efforts of a Catholic religious order to preserve knowledge. Another work, *Invasion of the Body Snatchers* (1955; made into a film classic in 1956), in a clear case of anticommunist paranoia, relates the story of ordinary people being replaced by look-alikes who operate as part of a collective body.

Anticommunist paranoia

Science fiction films of the period, with a few notable exceptions—such as *The Day the Earth Stood Still* (1951), *The War of the Worlds* (1953), and *Forbidden Planet* (1956)—tended to be cheaply produced, juvenile, formulaic films about alien invasions and monstrous mutants. (It was during this era that the Japanese produced numerous Godzilla movies.) In the genre's fiction, however, the American trio of Robert Heinlein, Isaac Asimov, and Ray Bradbury—later joined by Briton Arthur C. Clarke—enjoyed worldwide fame and unmatched popularity during the 1940s, '50s, and early '60s. In fact, Anglophone science fiction was dominant during the 1950s and '60s, though authors from other countries—such as the Polish *fantastyka* writer Stanisław Lem and the literary Italian Italo Calvino, with his *fantascienza*—also advanced the genre.

Change was also in the air in Soviet Russia. The political and cultural thaw that occurred during the rule of Nikita Khrushchev in the 1950s and the Russian-led dawn of the space age caused a dramatic upsurge in Soviet science fiction, including works by Ivan Yefremov, Kir Bulychev, and the renowned doyens of Russian-language science fiction, the brothers Arkady and Boris Strugatsky. A similar surge in Chinese science fiction accompanied the end of the Cultural Revolution (1966–76). In fact, at the start of the 21st century, China's main science fiction magazine claimed a readership of 500,000, dwarfing the circulation of any science fiction publication in the West.

In Britain and the United States, the editorial polemics of Michael Moorcock (associated for many years with *New Worlds* and its anthologies) and Harlan Ellison (*Dangerous Visions* [1967] and *Again, Dangerous Visions* [1972]) led a rebellious New Wave movement that facilitated the genre's move in fresh directions. Sporting a countercultural disregard of taboos (particularly with regard to morals and sexuality), a fascination with mind-altering drugs and Eastern religions, and an interest in experimental literary styles, the movement pushed the boundaries of traditional science fiction until the genre was almost unrecognizable. Most avant-garde experimentalism had vanished by the late 1970s, but by then the New Wave had vastly expanded the subgenre of "soft" science fiction. ("Soft" SF is typically more concerned with exploring social aspects of the near future and of "inner space," while "hard" science fiction features technology-for-technology's-sake).

New Wave

SF cinema and TV. In contrast to earlier decades, traditional science fiction of the late 1960s and early '70s reached unprecedented popularity on television and in film. American SF television series, such as *Star Trek* (1966–69; founded by Gene Roddenberry), may have primed film producers and audiences alike for cinema adaptations of "serious" science fiction. *Fahrenheit 451* (1966), *2001: A Space Odyssey* (1968), and *Charly* (1968)—based on works by Bradbury, Clarke, and Daniel Keyes, respectively—earned critical praise and attracted a growing number of directors and actors to the genre. If any doubt remained about the commercial viability of SF cinema, the blockbuster movies *Star Wars* (1977), *Close Encounters of the Third Kind* (1977), and *E.T.: The Extra-Terrestrial* (1982) proved that science fiction had finally moved beyond its drive-in B-film status. In fact, U.S. box-office receipts for science fiction, fantasy, and horror films jumped from 5 percent in 1971 to nearly 50 percent by 1982; although the share fell somewhat in subsequent years, science fiction continued to be one of the most important Hollywood movie formats.

Ridley Scott's film *Blade Runner* (1982), based on Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1968), prefigured the 1980s phenomenon known as cyberpunk. It combined a fascination for cybernetics (the science of communication and control theory, especially with regard to the human nervous system and brain) with a "punk," or alienated, social consciousness, thus melding elements of soft and hard science fiction. William Gibson in *Neuromancer* (1984) coined the word *cyberspace* to describe a computer-mediated virtual world into which humans plugged their brains. Other works of this subgenre include John Shirley's *City Come A-Walkin'* (1980), Bruce Sterling's *Schismatrix* (1985), and Lewis Shiner's *Deserted Cities of the Heart* (1988).

The spectacular nature of science fiction's thematics played very strongly to Hollywood's technical advantages over rival cinemas in Europe, Japan, Hong Kong, and Mumbai (Bombay). After the 1970s the American SF film with its state-of-the-art special effects became science fiction's public face. Science fiction films such as the Terminator series (1984, 1991, 2003), the Alien series (1979, 1986, 1992, 1997, 2004), and the Jurassic Park series (1993, 1997, 2001) became major money earners worldwide.

Heroic fantasy, which had remained a minority taste in Britain and elsewhere for many decades, captivated a new generation and emerged in the 1990s as a dominant subgenre known to devotees as "sword and sorcery." One indication of the changing commercial reality was the 1992 reorganization of SF's largest professional association, the Science Fiction Writers of America, into the Science Fiction and Fantasy Writers of America, Inc. Undreamed-of book sales of such fantasy works as J.K. Rowling's *Harry Potter and the Philosopher's Stone* (1997; also published as *Harry Potter and the Sorcerer's Stone*) and succeeding volumes brought wildly successful film adaptations of the Harry Potter books (2001, 2002, 2004, and on) as well as of J.R.R. Tolkien's *Lord of the Rings* (2001, 2002, and 2003).

Sword and sorcery

Major science fiction themes

UTOPIAS AND DYSTOPIAS

Sir Thomas More's satire *Utopia* (1516)—the title is based on a pun of the Greek words *eutopia* ("good place") and *outopia* ("no place")—shed an analytic light on 16th-century England along rational, humanistic lines. *Utopia* portrayed an ideal society in a hypothetical "no-place" so that More would be perceived as undertaking a thought experiment, giving no direct offense to established interests.

Since More's time, utopias have been attractive primarily to fringe political thinkers who have little practical redress within the power structures of the day. Under these conditions, a published thought experiment that airs hidden discontents can strike with revelatory force and find a broad popular response.

Utopias can be extravagant castles-in-the-air, nostalgic Shangri-Las, provocative satires, and rank political tracts thinly disguised as novels. Society's esteem for utopian thinking has fluctuated with the times. The failure of Soviet communism caused an immense archive of utopian work to shift catastrophically in value from sober social engineering to dusty irrelevancy. The line between reforming insight and political crankdom is often thin.

Utopias thrived amid the 19th century's infatuation with scientific progress. Many philosophers—Karl Marx included—thought that historical forces and the steady accumulation of rational knowledge would someday yield an "end state" for history. According to this way of thinking, the thoughtful futurist needed only to spot and nurture tomorrow's dominant progressive trends and kill off the feudal superstitions of false consciousness; then social perfection would arrive as surely as the ticking of a clock.

Fictional successes along this line included Edward Bellamy's *Looking Backward* (1888), in which a Bostonian awakes from a mystical sleep in the year 2000 to find industry nationalized, equal distribution of wealth to all citizens, and class divisions eradicated—a process that Bellamy called Nationalism. Bellamy Nationalist clubs sprang up nationwide to discuss his ideals, and the Nationalists were represented at the 1891 Populist Party convention; socialist leader Eugene V. Debs adopted many of the tenets of the Nationalist program. William Morris, who was appalled by Bellamy's depiction of a rational, bureaucratized industrial state, countered with *News from Nowhere*, a British vision of a pastoral utopia.

German politician Walther Rathenau wrote technological utopias, *Von kommenden Dingen* (1917; *In Days to Come*) and *Der neue Staat* (1919; *The New Society*), in which he rejected nationalized industries in favour of worker participation in management; in the turbulence of Weimar society, he was assassinated by anti-Semitic nationalists.

H.G. Wells became a particularly ardent and tireless socialist campaigner. In works such as *A Modern Utopia*

Socialist utopias

(1905), *Men Like Gods* (1923), *The Open Conspiracy: Blue Prints for a World Revolution* (1928), and *The Shape of Things to Come* (1933), he foresaw a rationalized, technocratic society. Yet Wells lived long enough to see the atomic bomb, and his last essay, "Mind at the End of Its Tether" (1945), darkly prophesied extinction for the human race, which, in his later opinion, lacked the creative flexibility to control its own affairs.

In B.F. Skinner's *Walden Two* (1948), rewards and punishments are employed to condition the members of a small communal society. In *Walden Two Revisited* (1976), Skinner was more explicit: "Russia after fifty years is not a model we wish to emulate. China may be closer to the solutions I have been talking about, but a communist revolution in America is hard to imagine."

Technocratic utopias like those envisioned by Wells and Skinner have a serious conceptual difficulty: where, how, and why is the process of "improvement" to stop? It is hard to champion "progress" by depicting a world in which further progress is impossible. This paradox does not apply to the pastoral utopia, which turns its back on technology to seek a timeless world of stability and peace. The pastoral utopia generally functions as an imaginary refuge from the technological forces that are so visibly warping the author's real-world landscape. Pastorals tend to be quiet, thoughtful village retreats devoid of smokestacks, newspapers, bank loans, and annoying traffic jams. Major works in this vein include Morris' *News from Nowhere*, Samuel Butler's satiric *Erewhon* (1872), James Hilton's *Lost Horizon* (1933), Aldous Huxley's psychedelic *Island* (1962), and Ernest Callenbach's green postindustrial *Ecotopia* (1975).

Pastoral utopias

Ursula K. Le Guin's *The Dispossessed* (1974) depicts an anarchist state striving to fulfill its own ideals, but, like most modern SF utopias, it emphasizes ambiguity rather than claiming that history is on the author's side. Kim Stanley Robinson's Martian Trilogy—*Red Mars* (1992), *Green Mars* (1994), and *Blue Mars* (1996)—describes planetary settlers creating an idealist pioneer society under Martian physical conditions.

A central difficulty of utopian fiction is the lack of dramatic conflict; a state of perfection is inherently uneventful. The counter to utopia is dystopia, in which hopes for betterment are replaced by electrifying fears of the ugly consequences of present-day behaviour. Utopias tend to have a placid gloss of phony benevolence, while dystopias display a somewhat satanic thunder.

Utopias commonly feature "moderns" undergoing a conversion experience to the utopian mind-set—after which all action stops. In dystopias a character representing moderns is excitingly chased down, persecuted, degraded, and commonly killed. In Huxley's *Brave New World*, an intellectual dissident is singled out and exiled by fatuous world rulers anxious to preserve their numbing status quo. George Orwell's hellish *Nineteen Eighty-four* stopped the march of history in its tracks with its image of the future as "a boot stamping on a human face—forever." Terry Gilliam's satiric film *Brazil* (1985) veers between pathos and absurdity with its bizarre blend of Orwell's dystopian vision of the future and Kafkaesque elements.

E.M. Forster's much-anthologized story "The Machine Stops" (1909) was written as a counterblast to Wellsian technical optimism. The story depicts a soulless, push-button, heavily networked world. The sudden collapse of Forster's dystopia supplies motive force to the plot—a scheme so common in science fiction that it is known as the "house-of-cards" plot.

In Norman Spinrad's black comedy *The Iron Dream* (1972), a frustrated Adolf Hitler immigrates and becomes an American pulp SF novelist, to weirdly convincing effect. Whether pleasant or sinister, heavenly or apocalyptic, utopias and dystopias share a sublime sense of ahistoricality. All solutions are necessarily final solutions, and the triumph, or calamity, would surely last at least a thousand years.

ALTERNATIVE SOCIETIES

If one abandons the odd notion that the passage of time must make things worse or better, the spectrum of possibility expands dramatically. Science fiction writers have spent much effort conceiving societies that are neither perfect nor horrific but excitingly different, alien to human experience. Robert Heinlein's greatest popular success, the novel *Stranger in a Strange Land* (1961), paints the fate of a prophet and social reformer who was raised by Martians. A Martian human has no earthly shibboleths, so the story's weird hero cuts briskly through almost every pious human custom relating to sex, death, religion, and money. For obvious reasons, Heinlein's work was a counterculture icon in the 1960s.

Many SF writers, like Heinlein, took particular pleasure in upsetting the most basic tenets of the human condition. John Varley's *The Ophiuchi Hotline* (1977) is an archive of methods to shatter old human verities: characters die and are reborn as clones, change sex with ease and alacrity, make backup tapes of their personalities, and undergo drastic acts of surgery—all in a space-dwelling society that accepts such things as normal.

William Gibson's *Neuromancer*, mentioned above, was widely noted for its intense depiction of a postnational world order ruled by feudal global corporations. Artificial intelligences, owned by the wealthy few, are hugely powerful entities, yet they pass almost unheeded over a seething, fractured society of outlaw geneticists, information criminals, colourful street gangs, and orbiting Rastafarians.

Neuromancer

In Neal Stephenson's *Snow Crash* (1992), a future globalized society has abandoned conventional land-based government and reformed itself along the lines of electronic cults and mobile interest groups. The Mafia delivers pizza, the CIA is a for-profit organization, Hong Kong is a global franchise of capitalist Chinatowns, and life online is often of more consequence than real life.

SEX AND GENDER

Because it is difficult to legislate relations between the sexes by conventional political reform, and because works of fiction can present a multiplicity of new arrangements, science fiction has had a particular affinity for feminism, and the attraction was mutual. In *Mizora* (1890), Mary Bradley Lane presented an early feminist utopia, and Charlotte Perkins Gilman in *Herland* (1915) imagined a society of women who reproduce by parthenogenesis.

The subject also interested some male authors. Theodore Sturgeon's *Venus Plus X* (1960) examines the limits of gender in a world where sexuality and reproduction are surgical add-ons. One of the more thoughtful explorations of the theme is Ursula K. Le Guin's *The Left Hand of Darkness* (1969), which posits a human society on a distant planet where humans have no sexual identity but become sexual beings for a brief period once a month; each can become either male or female during this time. Le Guin works out the consequences of this sort of arrangement in meticulous anthropological detail and creates a revelatory tour de force.

Because science fiction is by nature receptive to technical solutions to all sorts of issues, including gender, readers embraced Shulamith Firestone's feminist tract *The Dialectic of Sex: The Case for a Feminist Revolution* (1970); though the book was not written with a science fiction audience in mind, it nevertheless declared that women could never be free of oppression until the physical acts of childbearing and child rearing were industrialized. The influence of Firestone's book could be seen in works such as Marge Piercy's *Woman on the Edge of Time* (1976) and Suzy McKee Charnas' *Motherlines* (1978).

Although feminist SF tended to hope for gender justice, a powerful dystopian school of feminist science fiction suggested that relationships between men and women might slide from poor to downright catastrophic. Nazi cults of crazed masculinity haunt Katharine Burdekin's *Swastika Night* (1937). Joanna Russ's *The Female Man* (1975) suggests through its title that "femininity" is a weird condition forced on one by oppressors. Even Russ's feminist classic paled by comparison with Margaret Atwood's evocative dystopian misogyny in *The Handmaid's Tale* (1985). Drawn from dark contemporary trends, the bitter world of *The Handmaid's Tale* is ruled by a repressive American religious regime. This dystopia finally collapses from its own hostility to women—to be succeeded by yet another his-

The Handmaid's Tale

torical epoch. In this sense, *The Handmaid's Tale* makes an intellectual peace with historical process and transcends the customary limits of utopias and dystopias.

ALIEN ENCOUNTERS

Since human beings are the only known form of fully sentient life, any encounter with nonhuman intelligence is necessarily speculative. Writers in the 17th and 18th centuries produced many tales of travel to and from other inhabited worlds, but works such as Voltaire's *Micromégas* did not depict Saturnians as alien beings; they were men, though of Saturn-sized proportions.

A fuller knowledge of natural history enabled writers to imagine that life on other worlds might develop differently from life on Earth. In 1864 the astronomer and science popularizer Camille Flammarion published *Les Mondes imaginaires et les mondes réels* ("Imaginary Worlds and Real Worlds"), depicting otherworldly forms of life that could evolve within alien environments. This Gallic conceptual breakthrough was first exploited in fiction by J.H. Rosny Aîné, whose short story "Les Xipéhuz" (1887) describes an evolutionary war of extermination between prehistoric humans and a menacing crystal-based life-form.

Aliens were thus first conceived as Darwinian competitors with mankind, a scheme worked out in spooky Huxleyan detail by H.G. Wells, whose slimy, bloodsucking Martians possessed intellects "vast, and cool, and unsympathetic." Wells's *The War of the Worlds* (1898) was all the more successful for its implication that the highly advanced British Empire was finally experiencing from the other side the gunboat diplomacy that it had meted out to others. In 1938 Orson Welles's radio adaptation of *The War of the Worlds* was mistaken by the gullible for actual news reportage of marauding Martians sacking and looting New Jersey. The episode provoked an attack of mass panic, making it perhaps the most famous radio drama of all time.

Wells's *The First Men in the Moon* (1901) boosted antlike aliens into a sinister lunar analog for human society. The spate of alien invasion stories that followed were often strident in tone and genocidal in their predictions of coming doom. The "bug-eyed monster" became a staple of science fiction. Stanley G. Weinbaum won immediate and lasting acclaim with his more sophisticated approach in "A Martian Odyssey" (1934), which presented aliens whose behaviour, though whimsical, harmless, and colourful, was profoundly inexplicable to human mentality. In Raymond Z. Gallun's "Old Faithful" (1934), the Martians tend to be quite decent sorts.

Star Maker

Authors of "serious" literature, such as Olaf Stapledon, also dealt with alien life-forms. His *Star Maker* (1937) follows an Englishman whose disembodied mind travels across space and time, observing aliens as metaphysical actors in a fiery cosmic drama remote from all human concern, and encounters the creator of the universe (Star Maker). This critically acclaimed book is more a philosophical treatise on science, human nature, and God than a traditional novel. Stapledon's descriptions and social-philosophical discourses on galactic empires, symbiotic alien life-forms, genetic engineering, ecology, and overpopulation inspired a number of SF writers, including Arthur C. Clarke, during the 1940s and '50s.

As dramatic actors within a narrative, aliens pose unique difficulties. If too humanlike, they are of little use; if genuinely alien, they defy the fictional conventionalities of motive, conflict, and plot. In Stanisław Lem's *Solaris* (1961; filmed 1972, 2002), the sentience on an alien planet is so metaphysically distant from humanity that it causes its cosmonaut investigators to hallucinate and collapse. The Solaris alien is a permanent enigma, completely unframable by any human thought process. Hal Clement's *Mission of Gravity* (1954) was a tour de force in that its hero is a tiny, intelligent, centipede-like creature who breathes poison gas in the crushing gravity of an alien world. This description alone makes it clear just how difficult imagining the alien can be. As a result, science fiction writers often centred their energies on a first contact with aliens, such as those found in Steven Spielberg's film *Close Encounters of the Third Kind* (1977). In the "first contact" narrative, one can enjoy the novel thrill of alienness without having to confront the implications of everyday interactions with aliens.

Alien-invasion motifs persist in science fiction, as in the film *Alien* (1979) with its ruthless, parasitic monsters. Yet a distinct and growing trend within science fiction depicted aliens as coworkers, science officers, technical specialists, sidekicks, and even love interests. Two of the most prominent examples of this come from the various television shows, films, and novels based on the worlds of *Star Trek* and *Alien Nation*. It also became increasingly common for human characters to have undergone such extensive warping and mutation—as in Paul Di Fillipo's *Ribofunk* (1996)—that they themselves are as exotic as aliens.

Aliens are supposed evolutionary products of life on different worlds, while intelligent robots are supposed mechanical, industrial creations. Robots and aliens therefore serve similar thematic purposes for science fiction. The first robots were introduced by Czech dramatist Karel Čapek as characters in his play *R.U.R.* (1921). In a rather standard alien-menace maneuver, Čapek's robots outcompete humanity within the new milieu of industrial mass production and attempt to exterminate the human race.

Robots remain primarily theatrical inventions, but they are central figures in science fiction thought experiments intended to provoke debate about humanity's place within a technological environment. Isaac Asimov, for example, devoted much effort to creating an ethical system for humans and robots. Asimov's famous Three Laws of Robotics are as follows: "(1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

Asimov's Three Laws of Robotics

Asimov was able to derive an entertaining set of novels and stories from these three premises—even though his imaginary laws have never been used for the control of any real-world robot. Quite to the contrary, 21st-century robotics are probably best represented by semiautonomous military devices such as the cruise missile, specifically designed to blow itself up as it reaches its target and to do considerable damage.

The robot as a reflection of humanity received a classic outing in Lester del Rey's short story "Helen O'Loy" (1938). Helen was not the first female robot—her famous predecessor is the sinister celluloid robot Maria from the aforementioned film *Metropolis* (1927). Helen, by contrast, somehow establishes her womanhood by marrying her inventor and then sacrificing her own mechanical life upon her husband's death. Male robots, in the hands of authors such as Tanith Lee (*The Silver Metal Lover*, 1981) and Marge Piercy (*He, She, and It*, 1991), became distorted images of human men.

Humanoid robots, or androids, remain the photogenic darlings of SF cinema, appearing in a host of productions, including *Westworld* (1973), *The Stepford Wives* (1975, 2004), *Star Wars* (1977), *Bicentennial Man* (1999), *Artificial Intelligence: A.I.* (2001), and *I, Robot* (2004).

SPACE TRAVEL

Flight into outer space is the classic SF theme. Verne's pioneering *De la terre à la lune* (1865; *From the Earth to the Moon*) was the first fiction to treat space travel as a coherent engineering problem—to recognize explicitly that gravity would cease, that there could be no air, and so forth. Because Verne found no plausible way to land his cannon-fired passengers on the lunar surface, they merely whiz by the Moon at close range, cataloging craters in a geographic ecstasy. At the conceptual dawn of space travel, it was enough just to be up there, escaping earthly bonds to revel in sheer extraterrestrial possibility. Given that Georges Méliès filmed a fictional trip to the Moon with his pioneering camera in 1902, SF cinema is as old as cinema itself.

A certain disenchantment with this theme necessarily set in after the actual Moon landing in 1969, for human life in outer space proved less than heavenly. Far from swashbucklers, astronauts and cosmonauts were highly trained

(From left) "Bones" McCoy (DeForest Kelley), Captain Kirk (William Shatner), and Spock (Leonard Nimoy) in the transporter room; from the American television series *Star Trek* (1966–69).

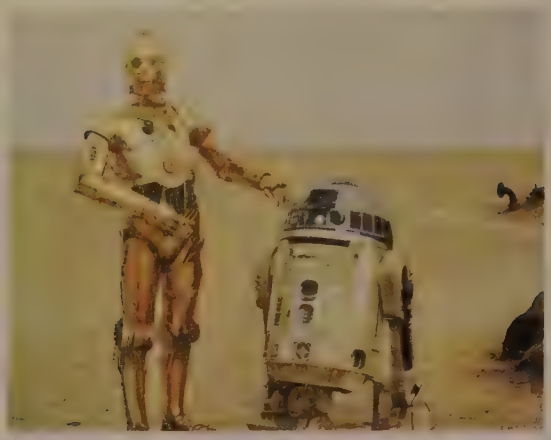
Movie poster depicting Queen Aelita (Yuliya Solntseva) in the Russian silent-film classic *Aelita* (1924).

Dust jacket for the first American edition of *The Handmaid's Tale* by Margaret Atwood, illustration by Fred Marcellino, published by Houghton Mifflin Company, 1986.

Movie poster for *Godzilla, King of the Monsters!* (1956).

The inventor C.A. Rotwang (Rudolf Klein-Rogge) and his robotic invention (Brigitte Helm) in Fritz Lang's silent-film classic *Metropolis* (1927).

(Top left) Paramount Pictures; photograph, Everett Collection. (top right, bottom right) Hershenson-Allen Archive (centre left) copyright Houghton Mifflin Company. Between the Covers Rare Books, Inc., Merchantville, New Jersey (bottom left) from a private collection



C-3PO (left) and R2-D2, played by Anthony Daniels and Kenny Baker, scene-stealing robots in the film *Star Wars* (1977).

Dr. Who, played by Tom Baker, in the episode "State of Decay" (1980), from the British television series *Dr. Who*.

(Left) Gort (left) and Klaatu, played by Lock Martin and Michael Rennie, in the film *The Day the Earth Stood Still* (1951).

(Right) David the robot (foreground), played by Haley Joel Osment, the subject of the film *Artificial Intelligence: A.I.* (2001).

(Below) Mumakil and their riders at the battle of Pelennor Fields, from the film *The Lord of the Rings: The Return of the King* (2003).

technicians whose primary motive was to preserve their hardware. They grappled with strict limits in fuel, power, water, oxygen, and privacy, along with cramped personal quarters—a life more akin to submarine service than to a romantic flight aboard a luxury starship.

The SF works that treat space travel with nuts-and-bolts realism are a minority taste. Science fiction far more commonly omits the unromantic aspects of space travel, especially through one of the genre's commonest stage devices, the "faster-than-light drive," or "warp drive." Although a warp drive is no more technically plausible than lifelike androids, it is a necessity for the alien-planet adventure story. Science fiction writers cheerfully sacrifice the realities of astrophysics in the service of imaginary worlds.

Space opera

Much creative energy has been invested in "space opera," science fiction at its most romantic. The space opera is an action-adventure, commonly of galactic scale, of which the film cycle *Star Wars* (1977, 1980, 1983, 1999, 2002) is the best-known exemplar. It presents a unique type of "widescreen baroque," with all the riches of pulp fiction in a single package. *Star Wars*, for example, offered not only advanced technology—presumably necessary to build the starships and orbiting battle stations—but also princesses, smugglers, robots, sword fights, mystical doctrines, levitating gurus, monsters, barroom brawls, heroes of dubious birth, elaborate chase scenes, and Gothic death traps.

Like the black-clad figures who move the props in Japanese nō theatre, the fantastic aspects of space opera are simply and gratefully accepted by its devotees. Writers of 20th-century space opera are among the most respected figures in science fiction. Their ranks include E.E. ("Doc") Smith, Edmond Hamilton, John W. Campbell, Jack Williamson, A.E. Van Vogt, Jack Vance, Anne McCaffrey, Lois McMaster Bujold, and C.J. Cherryh. Nor is space opera by any means moribund, for a particularly extravagant form of space opera is the signature of the New British Science Fiction, the first SF literary movement of the 21st century. Introverted postimperial insularity had long characterized British science fiction, but in the 21st century a cluster of writers—including Iain M. Banks, Stephen Baxter, Justina Robson, Peter F. Hamilton, Charles Stross, and Ken MacLeod—reengineered the universe in gaudy bursts of star-smashing neo-cosmology.

TIME TRAVEL

A complement to travel through space is travel through time. A prototype of the time-travel story is Charles Dickens' *A Christmas Carol* (1843). The story features the Ghost of Christmas Yet to Come, who is magically able to immerse the hapless Scrooge in the dire consequences of his own ungenerous actions. But for all their familiarity, Scrooge's time travels were mere ghostly dream mongering. The SF version of time travel arrived when H.G. Wells suggested in *The Time Machine* (1895) that the process might be done mechanically.

For a genre whose central issues involve processes of historical change, time travel is irresistibly attractive. For instance, time travel offers the edifying spectacle of "moderns" traveling into the past to remake the world closer to the heart's desire. Mark Twain's *A Connecticut Yankee in King Arthur's Court* (1889) contrasts industrial ingenuity with feudal romance, to darkly hilarious effect. L. Sprague de Camp's novel *Lest Darkness Fall* (1941) has an American archaeologist rescuing imperial Rome in its decline, an act the hero carries out with such luminous attention to techno-historical detail that it resembles a World Bank bailout of an underdeveloped country.

Time tourism, a distinct subgenre, is a perennial SF theme. It is exemplified in Ray Bradbury's "A Sound of Thunder" (1952), in which a tiny misstep by dinosaur hunters grimly affects the consequent course of history. In Robert Silverberg's *Up the Line* (1969), voyeuristic thrill seekers from the future infest the past.

Another variant on the time-travel theme involves physical objects that become displaced in time. C.M. Kornbluth's "The Little Black Bag" (1950) concerns a doctor's bag from the future. Warring groups of time travelers battle one another up and down the time streams in Poul Anderson's *Guardians of Time* (1960) and Fritz Leiber's *The Change War* (1978). Barrington J. Bayley's *Fall of Chronopolis* (1974) achieves the technicolour proportions of "time opera." In John Kessel's *Corrupting Dr. Nice* (1997), cynical exploiters from the future invade the past wholesale, kidnapping major historical figures and crassly employing them as underlings and talk-show hosts.

A one-way trip into the future is the staple of the suspended-animation story, the device behind the Buck Rogers stories and a host of consequent tales in which a hero of the present-day escapes the customary time-bound limits of human mortality. In Martin Amis' *Time's Arrow* (1991), the flow of time is entirely reversed, but life seems just as precarious as people solemnly march to a final end in their mothers' wombs.

Dr. Who

The long-lived British television series *Dr. Who* (1963–89) involved an eccentric time traveler whose exotic mode of transport was disguised as a common telephone booth. Periodically portrayed by different actors, Dr. Who exhibited a popularity so perennial that he indeed seemed timeless. The popularity of the notion can be seen in any number of time-travel films, including *The Time Machine* (1960 and 2002), *Slaughterhouse-Five* (1972), *Time Bandits* (1981), *Back to the Future* (1985), *Terminator* (1984), and *Twelve Monkeys* (1995).

ALTERNATE HISTORIES AND PARALLEL UNIVERSES

Stories centred on time-travel paradoxes developed as a separate school of science fiction. If a human being broke free from the conventional chains of causality, intriguing metaphysical puzzles ensued. The classic SF version of these puzzles is the challenge posed by a man who travels back in time and kills his own grandfather, thus ensuring that he, the time traveler, can never be born in the first place. Time-travel paradoxes were usually resolved as ingeniously as locked-room murder mysteries.

Murray Leinster's "Sidewise in Time" (1934) expanded the possibilities by suggesting a vast multiplicity of "histories," all occurring at the same "time." Under the scheme Leister proposed, one need not limit oneself to one past or one future but might travel between many alternate worlds existing in parallel. This new SF convention of a "multiverse" opened a vast potential canvas for fictional exploitation, with humanity's universe just one undistinguished universe among many.

Narratives set in the future offered at least some potential connection to the real world. By contrast, the "parallel universe" was entirely conjectural and hypothetical. Initially, readers found parallel worlds an amusing but inconsequential conceit, just as they had once found works set within the future. They soon realized, however, that the notion of uchronia (or "no-times") offered certain pleasures all its own, such as the ability to deploy actual historical figures as fictional characters. Well-known settings and events could be mutated and distorted at will.

Uchronic works

The passage of time had a complex, uchronic effect on science fiction itself. Despite the passing of the year 1984 itself, a number of concepts presented in *Nineteen Eighty-four*—such as omnipresent video surveillance—were not so far-fetched at the turn of the 21st century, and Orwell's political concerns remain painfully relevant. In addition to representing the uchronic effect of some works of science fiction, *Nineteen Eighty-four* is an excellent example of a uchronic novel; it is neither futuristic nor historical, existing in a peculiar uchronic netherworld. As time passes, growing numbers of SF classics fall into this conceptual category. It is a small step from this category to parallel worlds and alternate histories. Those concepts no longer seem abstract and improbable, but they have become part of the heritage of science fiction.

Even historical fiction has dealt with the "what if" posed by uchronias. In 1907 G.M. Trevelyan wrote an essay speculating on the consequences of a Napoleonic victory at Waterloo. Trevelyan's work inspired J.C. Squires's anthology *If It Had Happened Otherwise* (1931), in which such period worthies as Winston Churchill, André Maurois, and G.K. Chesterton speculated on counterfactual historical turning points. This was an intellectual parlour game of the type that science fiction liked to play.

Alternate histories existed well outside the customary

bounds of science fiction, such as Len Deighton's thriller *SS-GB* (1978), about the grim role of Nazi occupiers in Britain, and Vladimir Nabokov's involved and elegant *Ada* (1969). Alternate histories tend to cluster around particularly dramatic junctures of history, with World War II and the American Civil War as particular favourites. Some ventured farther out, postulating a global Roman Empire or a world in which dinosaurs avoided extinction.

The film *It's a Wonderful Life* (1946)—based on the story "The Greatest Gift" (1943) by Philip Van Doren Stern—is a perennial favourite. In the film, a man in despair learns that his life does matter when he sees that, without his presence, his hometown becomes an evil dystopia. It is an ultimate compliment to the individual when the universe rewrites itself around a fantasy of self-worth.

In some deep sense, all works of fiction must be alternate histories and parallel worlds, for their protagonists and described events do not in fact exist. As the tradition of fiction grew longer and deeper, presenting works ever more distant from the reader's cultural framework, readers seemed more willing to accept work that was radically detached from local truisms of time and space.

HIGH TECHNOLOGIES

Technological sublime

Leo Marx, author of the techno-social study *The Machine in the Garden* (1964), coined the useful term *technological sublime* to indicate a quasi-spiritual haze given off by any particularly visible and impressive technological advance. Science fiction dotes on the sublime, which ruptures the everyday and lifts the human spirit to the plateaus of high imagination. Common models of the technological sublime include railroads, photography, aviation, giant dams, rural electrification (a particular Soviet favourite), atomic power and atomic weapons, spaceflight, television, computers, virtual reality, and the "information superhighway." The most sublime of all technologies are, in reality, not technologies at all but rather technological concepts—time machines, interplanetary starships, and androids.

Humans quickly lose a sense of awe over the technological advancements that have been fully integrated into the fabric of everyday life. Technologies such as immunization, plumbing, recycling, and the birth control pill have had a profound cultural impact, but they are not considered sublime, nor are they generally subjects for science fiction. The reason for this is not directly related to the scientific principles involved or any inherent difficulties of the engineering. It is entirely a social judgment, with distinctly metaphysical overtones. Science fiction is one of the arenas in which these judgments are cast.

Spaceflight is one high technology to which science fiction has shown a passionate allegiance. For the most part, the space shuttle remained sublime even when it was three decades old and in its final years of operation. Were space shuttles as common as 747s, they would quickly lose their sublime effect.

Outer space and cyberspace—a science fiction term applied to computer networks and simulated spaces—are conceptual cousins, offering the same high-tech thrill through different instruments in different historical periods. Yet, as cybertechnology rapidly achieved mass acceptance and became commonplace in many parts of the 21st-century world, its SF allure faded. Science fiction therefore once again made tentative overtures to biotechnology, although a relationship has existed at least since Mary Shelley's *Frankenstein* was published. Unlike computers, biotechnology is deeply rooted in ancient and highly conservative pursuits such as medicine and agriculture. Social resistance to genetic alteration of crops, animals, and especially children is widespread.

The sheer novelty of computers masked their particular affinity for pornography, swindling, organized crime, and terrorist conspiracy until they were widely present in the home. By contrast, the potential social impact of cloning was easy to recognize and led to a spate of SF works, including Aldous Huxley's *Brave New World*, with its tank-born castes of workers. Czech "biopunk" stories of the 1980s used genetic parables to indict the moral warping of Czech society under Warsaw Pact oppression. Biologically altered "posthumans" became an SF staple. First visualized as menacing monsters or Nietzschean supermen, the genetically altered were increasingly seen as people with unconventional personal problems.

Although many of the technologies that were first envisioned by science fiction have become reality—and become mundane aspects of mainstream fictional works—scientific knowledge is growing exponentially, leaving plenty of room for further speculation about its future impact on society and individuals. It is hard to imagine any contemporary society's being fully immune to the prognosticating lure of science fiction.

BIBLIOGRAPHY

Origins of science fiction. PAUL K. ALKON, *Origins of Futuristic Fiction* (1987), examines 17th- through early 19th-century precursors of science fiction; and *Science Fiction Before 1900: Imagination Discovers Technology* (1994, reissued 2002), explores 19th-century scientific romance, a precursor of science fiction.

Encyclopaedias and general histories. BRIAN ALDISS and DAVID WINGROVE, *Trillion Year Spree: The History of Science Fiction* (1986, reprinted 2001), casts an objective yet interested eye on the SF world and its many unique customs and concepts. NEIL BARRON (ed.), *Anatomy of Wonder 4: A Critical Guide to Science Fiction*, 4th ed. (1995), is an exhaustive catalog of thousands of works of science fiction. DAVID PRINGLE (ed.), *The Ultimate Guide to Science Fiction: An A–Z of Science-Fiction Books by Title*, 2nd ed. (1995), by an influential British SF magazine editor, contains reviews of some 3,000 titles. JOHN CLUTE and PETER NICHOLLS (eds.), *The Encyclopedia of Science Fiction*, 2nd ed. (1999), is a work of remarkable scholastic rigour. JOHN CLUTE and JOHN GRANT (eds.), *The Encyclopedia of Fantasy* (1997, reissued 1999), explores the murkier byways of genre fantasy with illuminating results. BRIAN ASH (ed.), *The Visual Encyclopedia of Science Fiction* (1977), presents a great variety of the garish graphics typical of the genre.

Memoirs and culture. ERIC LEIF DAVIN, *Pioneers of Wonder: Conversations with the Founders of Science Fiction* (1999), contains interviews with science fiction's veterans of the 1920s and '30s regarding the largely forgotten world of prewar pulp fiction. FREDERIK POHL, *The Way the Future Was: A Memoir* (1978, reissued 1983), is one of the best and most deeply felt biographies of a hard-core SF professional. SAM MOSKOWITZ, *The Immortal Storm* (1954, reissued 1974), details the passionate allegiances and snarled feuds of science fiction's many amateur devotees; his *Explorers of the Infinite: Shapers of Science Fiction* (1963, reprinted 1974) profiles some of the movers and shakers of the field; and his *Seekers of Tomorrow* (1966, reprinted 1974) offers insight into the SF world at mid-century. HARRY WARNER, JR., *All Our Yesterdays: An Informal History of Science Fiction Fandom in the Forties* (1969), by a beloved figure in SF fandom, covers the SF scene.

Criticism. SAM J. LUNDWALL, *Science Fiction: What It's All About* (1971), presents the opinions of a Swedish writer. WILLIAM ATHELING, JR. (JAMES BLISH), *The Issue at Hand*, 2nd ed. (1974), collects Blish's early SF criticism. DARKO SUVIN, *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre* (1979), presents an analysis and definition of the SF genre. DAVID G. HARTWELL, *Age of Wonders: Exploring the World of Science Fiction* (1984, reissued 1996), is an opinionated look at science fiction by a noted American SF critic and editor. STANISŁAW LEM, *Microworlds: Writings on Science Fiction and Fantasy*, ed. by FRANZ ROTTENSTEINER (1984, reissued 1991), brings profound analytic brilliance to bear on the craft of science fiction, scattering wounded American SF writers right and left. DAMON KNIGHT, *In Search of Wonder*, 3rd ed. enlarged and extended (1996), collects the scathingly funny criticism of SF by a gifted editor and expert short-story writer.

Non-Western science fiction. DINGBO WU and PATRICK D. MURPHY (eds.), *Science Fiction from China* (1989), although having plots that may seem overly familiar to Western SF readers, has well-crafted stories. Noteworthy is the inclusion of an excellent history of Chinese science fiction, including its repression during the Cultural Revolution. JOHN L. APOSTOLOU and MARTIN H. GREENBERG (eds.), *The Best Japanese Science Fiction Stories* (1989, reissued 1997), although quite uneven in quality, contains mostly allegorical tales that offer the Western reader of science fiction a decidedly different experience. (Br.St.)

Diverse materials and techniques

Jade horse head, Chinese, Han dynasty (206 BC–AD 220). In the Victoria and Albert Museum, London. Height 19 cm.

Jaina pottery figurine, late classic Maya style, from the state of Campeche, Mexico. In the collection of Dumbarton Oaks, Washington, D.C. Height 15.5 cm.

"Virgin and Child," polychromed oak statue of the school of Auvergne, France, 12th century. In the Metropolitan Museum of Art, New York. 78.7 × 32.4 cm.

"Isaac, Jacob, and Esau," gilded bronze relief panel from the east doors ("Gates of Paradise") of the baptistery in Florence, by Lorenzo Ghiberti, 1425–52. 79.4 cm square.

"The Ecstasy of St. Teresa," marble and gilded bronze niche sculpture by Gian Lorenzo Bernini, 1645–52. In the Coronaro Chapel, Sta. Maria della Vittoria, Rome. Lifesize.

Plate 1 By courtesy of (top left) the Victoria and Albert Museum, London; (centre) Dumbarton Oaks, Washington, D.C.; (top right) the Metropolitan Museum of Art, New York, gift of J. Pierpont Morgan, 1916, photographs; (bottom left and right) SCALA—Art Resource/EB Inc

Painted wood male figure standing on a fish, fantastic cult image from Northern New Ireland, Melanesia. In a private collection. Height 1.53 m.

Diverse kinds of representational and nonrepresentational sculpture

"Cubi XVII," non-objective sculpture by David Smith, stainless steel, 1963. In the Dallas Museum of Fine Arts, Texas. Height 2.74 m.

"Development of a Bottle in Space," nonrepresentational sculpture by Umberto Boccioni, silvered bronze, 1912. In the Museum of Modern Art, New York. 38.1 × 32.7 × 59.7 cm.

"The Diner," representational environmental sculpture by George Segal, mixed media (plaster, wood, chrome, masonite, and formica), 1964–66. In the Walker Art Center, Minneapolis. 2.59 × 2.74 × 2.13 m.

the Eugene and Margaret McDermott Fund: (bottom) Collection

The Art of Sculpture

Sculpture is not a fixed term that applies to a permanently circumscribed category of objects or sets of activities. It is, rather, the name of an art that grows and changes and is continually extending the range of its activities and evolving new kinds of objects. The scope of the term is much wider in the second half of the 20th century than it was only two or three decades ago, and in the present fluid state of the visual arts nobody can predict what its future extensions are likely to be.

Certain features, which in previous centuries were considered essential to the art of sculpture, are not present in a great deal of modern sculpture and can no longer form part of its definition. One of the most important of these is representation. Before the 20th century, sculpture was considered a representational art; but its scope has now been extended to include nonrepresentational forms. It has long been accepted that the forms of such functional three-dimensional objects as furniture, pots, and buildings may be expressive and beautiful without being in any way representational; but it is only in the 20th century that nonfunctional, nonrepresentational, three-dimensional works of art have been produced.

Again, before the 20th century, sculpture was considered primarily an art of solid form, or mass. It is true that the negative elements of sculpture—the voids and hollows within and between its solid forms—have always been to some extent an integral part of its design, but their role has been a secondary one. In a great deal of modern sculpture, however, the focus of attention has shifted, and the spatial aspects have become dominant. Spatial sculpture is now a generally accepted branch of the art of sculpture.

It was also taken for granted in the sculpture of the past that its components were of a constant shape and size and did not move. With the recent development of kinetic sculpture, neither the immobility nor immutability of its form can any longer be considered essential to the art of sculpture.

Finally, 20th-century sculpture is not confined to the two traditional forming processes of carving and modelling or to such traditional natural materials as stone, metal, wood, ivory, bone, and clay. Because present-day sculptors use any materials and methods of manufacture that will serve their purposes, the art of sculpture can no longer be identified with any special materials or techniques.

Through all of these changes there is probably only one thing that has remained constant in the art of sculpture, and it is this that emerges as the central and abiding concern of sculptors: the art of sculpture is the branch of the visual arts that is especially concerned with the creation of expressive form in three dimensions.

Sculpture may be either in the round or in relief. A sculpture in the round is a separate, detached object in its own right, leading the same kind of independent existence in space as a human body or a chair. A relief does not have this kind of independence. It projects from and is attached to or is an integral part of something else that serves either as a background against which it is set or a matrix from which it emerges.

The actual three-dimensionality of sculpture in the round limits its scope in certain respects in comparison with the scope of painting. Sculpture cannot conjure up the illusion of space by purely optical means or invest its forms with atmosphere and light as painting can. It does have a kind of reality, a vivid physical presence that is denied to the pictorial arts. The forms of sculpture are tangible as well as visible, and they can appeal strongly and directly to both tactile and visual sensibilities. Blind people, even those who are congenitally blind, can produce and appreciate certain kinds of sculpture. It has, in fact, been argued by the 20th-century art critic Sir Herbert Read that sculpture should be regarded as primarily an art of touch and that the roots of sculptural sensibility can be traced to the pleasure one experiences in fondling things.

All three-dimensional forms are perceived as having an expressive character as well as purely geometric properties. They strike the observer as delicate, aggressive, flowing, taut, relaxed, dynamic, soft, and so on. By exploiting the expressive qualities of form, a sculptor is able to create images in which subject matter and expressiveness of form are mutually reinforcing. Such images go beyond the mere presentation of fact and communicate a wide range of subtle and powerful feelings.

The aesthetic raw material of sculpture is, so to speak, the whole realm of expressive three-dimensional form. A sculpture may draw upon what already exists in the endless variety of natural and man-made form, or it may be an art of pure invention. It has been used to express a vast range of human emotions and feelings from the most tender and delicate to the most violent and ecstatic.

All human beings, intimately involved from birth with the world of three-dimensional form, learn something of its structural and expressive properties and develop emotional responses to them. This combination of understanding and sensitive response, often called a sense of form, can be cultivated and refined. It is to this sense of form that the art of sculpture primarily appeals.

This article deals with the elements and principles of design; the materials, methods, techniques, and forms of sculpture; and its subject matter, imagery, symbolism, and uses. For the history of sculpture in the West, see SCULPTURE, THE HISTORY OF WESTERN. For treatments of sculpture as practiced in non-European cultures, see AFRICAN ARTS; AMERICAN PEOPLES, ARTS OF NATIVE; CENTRAL ASIAN ARTS; EAST ASIAN ARTS; EGYPTIAN ARTS AND ARCHITECTURE, ANCIENT; ISLĀMIC ARTS; OCEANIC ARTS; PREHISTORIC PEOPLES AND CULTURES; SOUTH ASIAN ARTS; SOUTHEAST ASIAN ARTS.

The article is divided into the following sections:

Elements and principles of sculptural design 43
- Elements of design 44
- Principles of design 45
- Relationships to other arts 46

Materials 46
- Primary 46
- Secondary 48

Methods and techniques 49
- The sculptor as designer and as craftsman 49
- Reproduction and surface-finishing techniques 52

Forms, subject matter, imagery, and symbolism of sculpture 53
- Sculpture in the round 54
- Relief sculpture 54
- Modern forms of sculpture 55
- Representational sculpture 56
- Nonrepresentational sculpture 58
- Decorative sculpture 58
- Symbolism 58

Uses of sculpture 59

Bibliography 59

Elements and principles of sculptural design

Ways in which space enters into the design of sculpture

The two most important elements of sculpture—mass and space—are, of course, separable only in thought. All sculpture is made of a material substance that has mass and exists in three-dimensional space. The mass of sculpture is thus the solid, material, space-occupying bulk that is contained within its surfaces. Space enters into the design of sculpture in three main ways: the material components

of the sculpture extend into or move through space; they may enclose or enfold space, thus creating hollows and voids within the sculpture; and they may relate one to another across space. Volume, surface, light and shade, and colour are supporting elements of sculpture.

ELEMENTS OF DESIGN

The amount of importance attached to either mass or space in the design of sculpture varies considerably. In Egyptian sculpture and in most of the sculpture of the 20th-century artist Constantin Brancusi, for example, mass is paramount, and most of the sculptor's thought has been devoted to shaping a lump of solid material. In 20th-century works by Antoine Pevsner or Naum Gabo, on the other hand, mass is reduced to a minimum, consisting only of transparent sheets of plastic or thin metal rods. The solid form of the components themselves is of little importance; their main function is to create movement through space and to enclose space. In works by such 20th-century sculptors as Henry Moore and Barbara Hepworth, the elements of space and mass are treated as more or less equal partners.

It is not possible to see the whole of a fully three-dimensional form at once. The observer can only see the whole of it if he turns it around or goes around it himself. For this reason it is sometimes mistakenly assumed that sculpture must be designed primarily to present a series of satisfactory projective views and that this multiplicity of views constitutes the main difference between sculpture and the pictorial arts, which present only one view of their subject. Such an attitude toward sculpture ignores the fact that it is possible to apprehend solid forms as volumes, to conceive an idea of them in the round from any one aspect. A great deal of sculpture is designed to be apprehended primarily as volume.

A single volume is the fundamental unit of three-dimensional solid form that can be conceived in the round. Some sculptures consist of only one volume, others are configurations of a number of volumes. The human figure is often treated by sculptors as a configuration of volumes, each of which corresponds to a major part of the body, such as the head, neck, thorax, and thigh.

Holes and cavities in sculpture, which are as carefully shaped as the solid forms and are of equal importance to the overall design, are sometimes referred to as negative volumes.

The surfaces of sculpture are in fact all that one actually sees. It is from their inflections that one makes inferences about the internal structure of the sculpture. A surface has, so to speak, two aspects: it contains and defines the internal structure of the masses of the sculpture, and it is the part of the sculpture that enters into relations with external space.

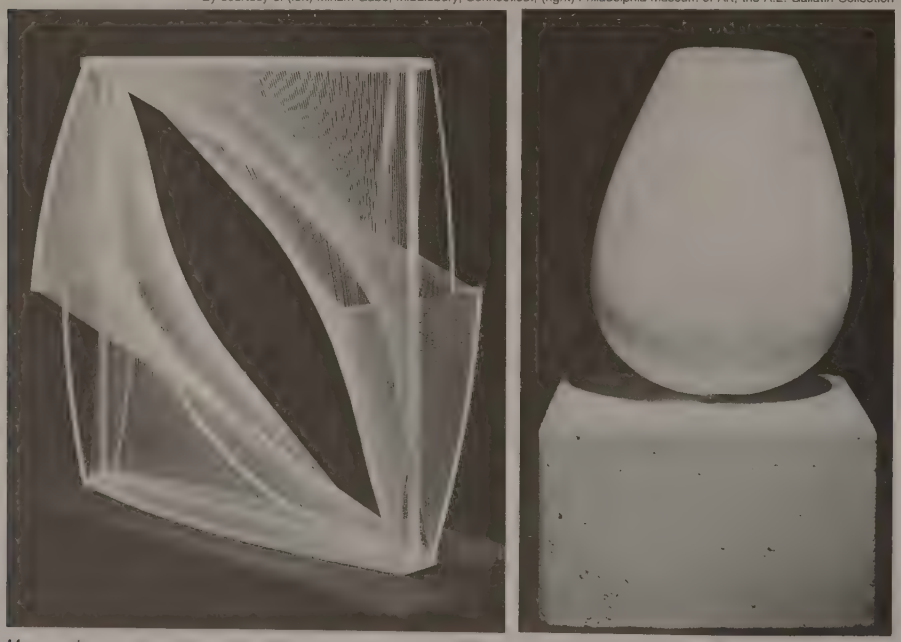
The expressive character of different kinds of surfaces is of the utmost importance in sculpture. Double-curved convex surfaces suggest fullness, containment, enclosure, the outward pressure of internal forces. In the aesthetics of Indian sculpture such surfaces have a special metaphysical significance. Representing the encroachment of space into the mass of the sculpture, concave surfaces suggest the action of external forces and are often indicative of collapse or erosion. Flat surfaces tend to convey a feeling of material hardness and rigidity; they are unbending or unyielding, unaffected by either internal or external pressures. Surfaces that are convex in one curvature and concave in the other can suggest the operation of internal pressures and at the same time a receptivity to the influence of external forces. They are associated with growth, with expansion into space.

The sculptor cannot, like the painter, create his own lighting effects within the work itself. The distribution of light and shade over the forms of his work depends upon the direction and intensity of light from external sources. Nevertheless, to some extent he can determine the kinds of effect this external light will have. If he knows where the work is to be sited, he can adapt it to the kind of light it is likely to receive. The brilliant overhead sunlight of Egypt and India demands a different treatment from the dim interior light of a northern medieval cathedral. Then again, it is possible to create effects of light and shade, or chiaroscuro, by cutting or modelling deep, shadow-catching hollows and prominent, highlighted ridges. Many late Gothic sculptors used light and shade as a powerful expressive feature of their work, aiming at a mysterious obscurity, with forms broken by shadow emerging from a dark background. Greek, Indian, and most Italian Renaissance sculptors shaped the forms of their work to receive light in a way that makes the whole work radiantly clear.

Chiaroscuro effects

The colouring of sculpture may be either natural or applied. In the recent past, sculptors became more aware than ever before of the inherent beauty of sculptural materials. Under the slogan of "truth to materials" many of

By courtesy of (left) Miriam Gabo, Middlebury, Connecticut, (right) Philadelphia Museum of Art, the A.E. Gallatin Collection



Mass and space.
(Left) "Linear Construction #1, Variation," Perspex plastic and nylon thread sculpture by Naum Gabo, 1942–43. In the Miriam Gabo Collection, Middlebury, Connecticut. 62 cm × 62 cm. (Right) "Torso of a Young Girl," onyx on a stone base by Constantin Brancusi, 1922. In the Philadelphia Museum of Art. Onyx height 34.9 cm; base height 17.1 cm.

them worked their materials in ways that exploited their natural properties, including colour and texture. More recently, however, there has been a growing tendency to use bright artificial colouring as an important element in the design of sculpture.

In the ancient world and during the Middle Ages almost all sculpture was artificially coloured, usually in a bold and decorative rather than a naturalistic manner. The sculptured portal of a cathedral, for example, would be coloured and gilded with all the brilliance of a contemporary illuminated manuscript. Combinations of differently coloured materials, such as the ivory and gold of some Greek sculpture, were not unknown before the 17th century; but the early Baroque sculptor Gian Lorenzo Bernini greatly extended the practice by combining variously coloured marbles with white marble and gilt bronze.

PRINCIPLES OF DESIGN

It is doubtful whether any principles of design are universal in the art of sculpture, for the principles that govern the organization of the elements of sculpture into expressive compositions differ from style to style. In fact, distinctions made among the major styles of sculpture are largely based on a recognition of differences in the principles of design that underlie them. Thus, the art historian Erwin Panofsky was attempting to define a difference of principle in the design of Romanesque and Gothic sculpture when he stated that the forms of Romanesque were conceived as projections from a plane outside themselves, while those of Gothic were conceived as being centred on an axis within themselves. The "principle of axiality" was considered by Panofsky to be "the essential principle of classical statuary," which Gothic had rediscovered.

The principles of sculptural design govern the approaches of sculptors to such fundamental matters as orientation, proportion, scale, articulation, and balance.

For conceiving and describing the orientation of the forms of sculpture in relation to each other, to a spectator, and to their surroundings, some kind of spatial scheme of reference is required. This is provided by a system of axes and planes of reference.

An axis is an imaginary centre line through a symmetrical or near symmetrical volume or group of volumes. Thus, all the main components of the human body have axes of their own, while an upright figure has a single vertical axis running through its entire length. Volumes may rotate or tilt on their axes.

Planes of reference are imaginary planes to which the movements, positions, and directions of volumes, axes, and surfaces may be referred. The principal planes of reference are the frontal, the horizontal, and the two profile planes.

The principles that govern the characteristic poses and spatial compositions of upright figures in different styles of sculpture are formulated with reference to axes and the four cardinal planes: for example, the principle of axiality already referred to; the principle of frontality, which governs the design of Archaic sculpture; the characteristic contrapposto (pose in which parts of the body, such as upper and lower, are twisted in opposite directions) of Michelangelo's figures; and in standing Greek sculpture of the Classical period the frequently used balanced "chiastic" pose (stance in which the body weight is taken principally on one leg, thereby creating a contrast of tension and relaxation between the opposite sides of a figure).

Proportional relations exist among linear dimensions, areas, and volumes and masses. All three types of proportion coexist and interact in sculpture, contributing to its expressiveness and beauty. Attitudes toward proportion differ considerably among sculptors. Some sculptors, both abstract and figurative, use mathematical systems of proportion; for example, the refinement and idealization of natural human proportions was a major preoccupation of Greek sculptors. Indian sculptors employed iconometric canons, or systems of carefully related proportions, that determined the proportions of all significant dimensions of the human figure. African and other tribal sculptors base the proportions of their figures on the subjective importance of the parts of the body. Unnatural proportions

Proportion.
Four figures possibly of the royal family of Judah, stone, 1145–50. Portail Royal of Chartres cathedral, France. Height approximately 2.50 m.
ND photo

Use of unnatural proportions

may be used for expressive purposes or to accommodate a sculpture to its surroundings. The elongation of the figures on the Portail Royal ("Royal Portal") of Chartres cathedral does both: it enhances their otherworldliness and also integrates them with the columnar architecture.

Sometimes it is necessary to adapt the proportions of sculpture to suit its position in relation to a viewer. A figure sited high on a building, for example, is usually made larger in its upper parts in order to counteract the effects of foreshortening. This should be allowed for when a sculpture intended for such a position is exhibited on eye level in a museum.

The scale of sculpture must sometimes be considered in relation to the scale of its surroundings. When it is one element in a larger complex, such as the facade of a building, it must be in scale with the rest. Another important consideration that sculptors must take into account when designing outdoor sculpture is the tendency of sculpture in the open air—particularly when viewed against the sky—

to appear less massive than it does in a studio. Because one tends to relate the scale of sculpture to one's own human physical dimensions, the emotional impact of a colossal figure and a small figurine are quite different.

In ancient and medieval sculpture the relative scale of the figures in a composition is often determined by their importance; *e.g.*, slaves are much smaller than kings or nobles. This is sometimes known as hierarchic scale.

The joining of one form to another may be accomplished in a variety of ways. In much of the work of the 19th-century French sculptor Auguste Rodin, there are no clear boundaries, and one form is merged with another in an impressionistic manner to create a continuously flowing surface. In works by the Greek sculptor Praxiteles, the forms are softly and subtly blended by means of smooth, blurred transitions. The volumes of Indian sculpture and the surface anatomy of male figures in the style of the Greek sculptor Polyclitus are sharply defined and clearly articulated. One of the main distinctions between the work of Italian and northern Renaissance sculptors lies in the Italians' preference for compositions made up of clearly articulated, distinct units of form and the tendency of the northern Europeans to subordinate the individual parts to the allover flow of the composition.

(Left) Collection, the Museum of Modern Art, New York, Mrs Simon Guggenheim Fund; (right) Alinari—Art Resource

Articulation.
(Left) Flowing surface exemplified by "St. John the Baptist Preaching," bronze sculpture by Auguste Rodin, 1878. In the Museum of Modern Art, New York City. Height 2 m. (Right) Delineated surface exemplified by "L'Idolino" bronze, Roman copy of a Greek sculpture in the style of Polyclitus, *c.* 440 BC. In the Museo Archeologico, Florence.

The balance, or equilibrium, of freestanding sculpture has three aspects. First, the sculpture must have actual physical stability. This can be achieved by natural balance—that is, by making the sculpture stable enough in itself to stand firmly—which is easy enough to do with a four-legged animal or a reclining figure but not with a standing figure or a tall, thin sculpture, which must be secured to a base. The second aspect of balance is compositional. The interaction of forces and the distribution of weight within a composition may produce a state of either dynamic or static equilibrium. The third aspect of balance applies only to sculpture that represents a living figure. A live human figure balances on two feet by making constant movements and muscular adjustments. Such an effect can be conveyed in sculpture by subtle displacements of form and suggestions of tension and relaxation.

RELATIONSHIPS TO OTHER ARTS

Sculpture has always been closely related to architecture through its role as architectural decoration and also at the level of design. Architecture, like sculpture, is concerned with three-dimensional form; and, although the central problem in the design of buildings is the organization of space rather than mass, there are styles of architecture that are effective largely through the quality and organization of their solid forms. Ancient styles of stone architecture, particularly Egyptian, Greek, and Mexican, tend to treat their components in a sculptural manner. Moreover, most buildings viewed from the outside are compositions of masses. The growth of spatial sculpture is so intimately related to the opening up and lightening of architecture, which the development of modern building technology has made possible, that many 20th-century sculptors can be said to treat their work in an architectural manner.

Sculpture and the pictorial arts

Some forms of relief sculpture approach very closely the pictorial arts of painting, drawing, engraving, and so on. And sculptures in the round that make use of chiaroscuro and that are conceived primarily as pictorial views rather than as compositions in the round are said to be "painterly"; for example, Bernini's "Ecstasy of St. Teresa" (Sta. Maria della Vittoria, Rome).

The borderlines between sculpture and pottery and the metalworking arts are not clear-cut, and many pottery and metal artifacts have every claim to be considered as sculpture. Today there is a growing affinity between the work of industrial designers and sculptors. Sculptural modelling techniques, and sometimes sculptors themselves, are often involved, for example, in the initial stages of the design of new automobile bodies.

The close relationships that exist between sculpture and the other visual arts are attested by the number of artists who have readily turned from one art to another; for example, Michelangelo, Bernini, Pisanello, Degas, and Picasso.

Materials

Any material that can be shaped in three dimensions can be used sculpturally. Certain materials, by virtue of their structural and aesthetic properties and their availability, have proved especially suitable. The most important of these are stone, wood, metal, clay, ivory, and plaster. There are also a number of materials of secondary importance and many that have only recently come into use.

PRIMARY

Throughout history, stone has been the principal material of monumental sculpture. There are practical reasons for this: many types of stone are highly resistant to the weather and therefore suitable for external use; stone is available in all parts of the world and can be obtained in large blocks; many stones have a fairly homogeneous texture and a uniform hardness that make them suitable for carving; stone has been the chief material used for the monumental architecture with which so much sculpture has been associated.

Stones belonging to all three main categories of rock formation have been used in sculpture. Igneous rocks, which are formed by the cooling of molten masses of mineral as they approach the Earth's surface, include granite, diorite, basalt, and obsidian. These are some of the hardest stones used for sculpture. Sedimentary rocks, which include sandstones and limestones, are formed from accumulated deposits of mineral and organic substances. Sandstones are agglomerations of particles of eroded stone held together by a cementing substance. Limestones are formed chiefly from the calcareous remains of organisms. Alabaster (gypsum), also a sedimentary rock, is a chemical deposit. Many varieties of sandstone and limestone, which vary greatly in quality and suitability for carving, are used for sculpture. Because of their method of formation, many sedimentary rocks have pronounced strata and are rich in fossils.

Metamorphic rocks result from changes brought about in the structure of sedimentary and igneous rocks by extreme pressure or heat. The most well-known metamorphic rocks used in sculpture are the marbles, which are recrystallized limestones. Italian Carrara marble, the best

known, was used by Roman and Renaissance sculptors, especially Michelangelo, and is still widely used. The best known varieties used by Greek sculptors, with whom marble was more popular than any other stone, are Pentelic—from which the Parthenon and its sculpture are made—and Parian.

Because stone is extremely heavy and lacks tensile strength, it is easily fractured if carved too thinly and not properly supported. A massive treatment without vulnerable projections, as in Egyptian and pre-Columbian American Indian sculpture, is therefore usually preferred. Some stones, however, can be treated more freely and openly; marble in particular has been treated by some European sculptors with almost the same freedom as bronze, but such displays of virtuosity are achieved by overcoming rather than submitting to the properties of the material itself.

Colours and textures of stone

The colours and textures of stone are among its most delightful properties. Some stones are fine-grained and can be carved with delicate detail and finished with a high polish; others are coarse-grained and demand a broader treatment. Pure white Carrara marble, which has a translucent quality, seems to glow and responds to light in a delicate, subtle manner. (These properties of marble were brilliantly exploited by 15th-century Italian sculptors such as Donatello and Desiderio da Settignano.) The colouring of granite is not uniform but has a salt-and-pepper quality and may glint with mica and quartz crystals. It may be predominantly black or white or a variety of grays, pinks, and reds. Sandstones vary in texture and are often warmly coloured in a range of buffs, pinks, and reds. Limestones vary greatly in colour, and the presence of fossils may add to the interest of their surfaces. A number of stones are richly variegated in colour by the irregular veining that runs through them.

Hardstones, or semiprecious stones, constitute a special group, which includes some of the most beautiful and decorative of all substances. The working of these stones, along with the working of more precious gemstones, is usually considered as part of the glyptic (gem carving or engraving), or lapidary, arts, but many artifacts produced from them can be considered small-scale sculpture. They are often harder to work than steel. First among the hardstones used for sculpture is jade, which was venerated by the ancient Chinese, who worked it, together with other hardstones, with extreme skill. It was also used sculpturally by Mayan and Mexican artists. Other important hardstones are rock crystal, rose quartz, amethyst, agate, and jasper.

The principal material of tribal sculpture in Africa, Oceania, and North America, wood has also been used by every great civilization; it was used extensively during the Middle Ages, for example, especially in Germany and central Europe. Among modern sculptors who have used wood for important works are Ernst Barlach, Ossip Zadkine, and Henry Moore.

Characteristics of wood

Both hardwoods and softwoods are used for sculpture. Some are close-grained, and they cut like cheese; others are open-grained and stringy. The fibrous structure of wood gives it considerable tensile strength, so that it may be carved thinly and with greater freedom than stone. For large or complex open compositions, a number of pieces of wood may be jointed. Wood is used mainly for indoor sculpture, for it is not as tough or durable as stone; changes of humidity and temperature may cause it to split, and it is subject to attack by insects and fungus. The grain of wood is one of its most attractive features, giving variety of pattern and texture to its surfaces. Its colours, too, are subtle and varied. In general, wood has a warmth that stone does not have, but it lacks the massive dignity and weight of stone.

The principal woods for sculpture are oak, mahogany, limewood, walnut, elm, pine, cedar, boxwood, pear, and ebony; but many others are also used. The sizes of wood available are limited by the sizes of trees; North American Indians, for example, could carve gigantic totem poles in pine, but boxwood is available only in small pieces.

In the 20th century, wood was being used by many sculptors as a medium for construction as well as for carving. Laminated timbers, chipboards, and timber in block and plank form can be glued, jointed, screwed, or bolted together, and given a variety of finishes.

Wherever metal technologies have been developed, metals have been used for sculpture. The amount of metal sculpture that has survived from the ancient world does not properly reflect the extent to which it was used, for vast quantities have been plundered and melted down. Countless Far Eastern and Greek metal sculptures have been lost in this way, as has almost all the goldwork of pre-Columbian American Indians.

The metal most used for sculpture is bronze, which is basically an alloy of copper and tin; but gold, silver, aluminum, copper, brass, lead, and iron have also been widely used. Most metals are extremely strong, hard, and durable, with a tensile strength that permits a much greater freedom of design than is possible in either stone or wood.

Advantages of bronze

A life-size bronze figure that is firmly attached to a base needs no support other than its own feet and may even be poised on one foot. Considerable attenuation of form is also possible without risk of fracture.

The colour, brilliant lustre, and reflectivity of metal surfaces have been highly valued and made full use of in sculpture although, since the Renaissance, artificial patinas have generally been preferred as finishes for bronze.

Metals can be worked in a variety of ways in order to produce sculpture. They can be cast—that is, melted and poured into molds; squeezed under pressure into dies, as in coin making; or worked directly—for example, by hammering, bending, cutting, welding, and repoussé (hammered or pressed in relief).

Important traditions of bronze sculpture are Greek, Roman, Indian (especially Cōḷa), African (Bini and Yoruba), Italian Renaissance, and Chinese. Gold was used to great effect for small-scale works in pre-Columbian America and medieval Europe. A fairly recent discovery, aluminum has been used a great deal by modern sculptors. Iron has not been used much as a casting material, but in recent years it has become a popular material for direct working by techniques similar to those of the blacksmith. Sheet metal is one of the principal materials used nowadays for constructional sculpture. Stainless steel in sheet form has been used effectively by the American sculptor David Smith.

Clay is one of the most common and easily obtainable of all materials. Used for modelling animal and human figures long before men discovered how to fire pots, it has been one of the sculptor's chief materials ever since.

Clay has four properties that account for its widespread use: when moist, it is one of the most plastic of all substances, easily modelled and capable of registering the most detailed impressions; when partially dried out to a leather-hard state or completely dried, it can be carved and scraped; when mixed with enough water, it becomes a creamy liquid known as slip, which may be poured into molds and allowed to dry; when fired to temperatures of between 700° and 1,400° C (1,300° and 2,600° F), it undergoes irreversible structural changes that make it permanently hard and extremely durable.

Uses of clay

Sculptors use clay as a material for working out ideas; for preliminary models that are subsequently cast in such materials as plaster, metal, and concrete or carved in stone; and for pottery sculpture.

Depending on the nature of the clay body itself and the temperature at which it is fired, a finished pottery product is said to be earthenware, which is opaque, relatively soft, and porous; stoneware, which is hard, nonporous, and more or less vitrified; or porcelain, which is fine-textured, vitrified, and translucent. All three types of pottery are used for sculpture. Sculpture made in low-fired clays, particularly buff and red clays, is known as terra-cotta (baked earth). This term is used inconsistently, however, and is often extended to cover all forms of pottery sculpture.

Unglazed clay bodies can be smooth or coarse in texture and may be coloured white, gray, buff, brown, pink, or red. Pottery sculpture can be decorated with any of the techniques invented by potters and coated with a variety of beautiful glazes.

Paleolithic sculptors produced relief and in-the-round work in unfired clay. The ancient Chinese, particularly

during the T'ang (618–907) and Sung (960–1279) dynasties, made superb pottery sculpture, including large-scale human figures. The best known Greek works are the intimate small-scale figures and groups from Tanagra. Mexican and Mayan sculptor-potters produced vigorous, directly modelled figures. During the Renaissance, pottery was used in Italy for major sculptural projects, including the large-scale glazed and coloured sculptures of Luca Della Robbia and his family, which are among the finest works in the medium. One of the most popular uses of the pottery medium has been for the manufacture of figurines—at Staffordshire, Meissen, and Sèvres, for example.

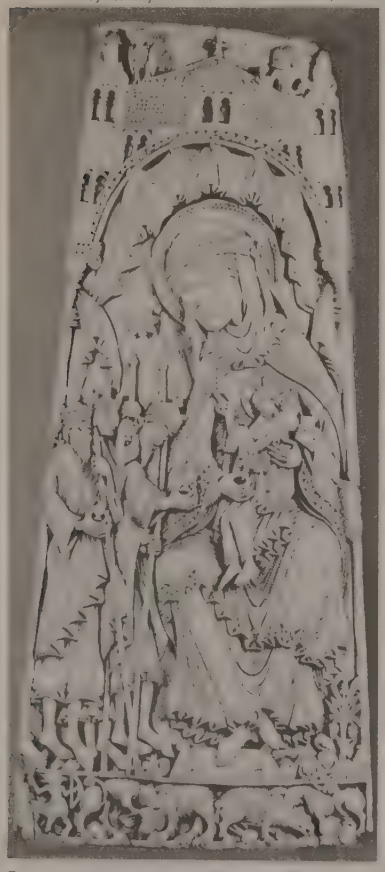
Sources of ivory

The main source of ivory is elephant tusks; but walrus, hippopotamus, narwhal (an Arctic aquatic animal), and, in Paleolithic times, mammoth tusks also were used for sculpture. Ivory is dense, hard, and difficult to work. Its colour is creamy white, which usually yellows with age; and it will take a high polish. A tusk may be sawed into panels for relief carving or into blocks for carving in the round; or the shape of the tusk itself may be used. The physical properties of the material invite the most delicate, detailed carving, and displays of virtuosity are common.

Ivory was used extensively in antiquity in the Middle and Far East and the Mediterranean. An almost unbroken Christian tradition of ivory carving reaches from Rome and Byzantium to the end of the Middle Ages. Throughout this time, ivory was used mainly in relief, often in conjunction with precious metals, enamels, and precious stones to produce the most splendid effects. Some of its main sculptural uses were for devotional diptychs, portable altars, book covers, retables (raised shelves above altars), caskets, and crucifixes. The Baroque period, too, is rich in ivories, especially in Germany. A fine tradition of ivory carving also existed in Benin, a former kingdom of West Africa.

Related to ivory, horn and bone have been used since Paleolithic times for small-scale sculpture. Reindeer horn and walrus tusks were two of the Eskimo carver's most important materials. One of the finest of all medieval "ivories" is a carving in whalebone, "The Adoration of the Magi" (Victoria and Albert Museum, London).

By courtesy of Victoria and Albert Museum, London



Bone carving.
"The Adoration of the Magi," whalebone, English, 11th–12th century. In the Victoria and Albert Museum, London. Height 37 cm.

Plaster of Paris (sulfate of lime) is especially useful for the production of molds, casts, and preliminary models. It was used by Egyptian and Greek sculptors as a casting medium and is today the most versatile material in the sculptor's workshop.

When mixed with water, plaster will in a short time recrystallize, or set—that is, become hard and inert—and its volume will increase slightly. When set, it is relatively fragile and lacking in character and is therefore of limited use for finished work. Plaster can be poured as a liquid, modelled directly when of a suitable consistency, or easily carved after it has set. Other materials can be added to it to retard its setting, to increase its hardness or resistance to heat, to change its colour, or to reinforce it.

The main sculptural use of plaster in the past was for molding and casting clay models as a stage in the production of cast metal sculpture. Many sculptors today omit the clay-modelling stage and model directly in plaster. As a mold material in the casting of concrete and fibreglass sculpture, plaster is widely used. It has great value as a material for reproducing existing sculpture; many museums, for example, use such casts for study purposes.

SECONDARY

Basically, concrete is a mixture of an aggregate (usually sand and small pieces of stone) bound together by cement. A variety of stones, such as crushed marble, granite chips, and gravel, can be used, each giving a different effect of colour and texture. Commercial cement is gray, white, or black; but it can be coloured by additives. The cement most widely used by sculptors is *ciment fondu,* which is extremely hard and quick setting. A recent invention—at least, in appropriate forms for sculpture—concrete is rapidly replacing stone for certain types of work. Because it is cheap, hard, tough, and durable, it is particularly suitable for large outdoor projects, especially decorative wall surfaces. With proper reinforcement it permits great freedom of design. And by using techniques similar to those of the building industry, sculptors are able to create works in concrete on a gigantic scale.

Advantages of concrete

When synthetic resins, especially polyesters, are reinforced with laminations of glass fibre, the result is a lightweight shell that is extremely strong, hard, and durable. It is usually known simply as fibreglass. After having been successfully used for car bodies, boat hulls, and the like, it has developed recently into an important material for sculpture. Because the material is visually unattractive in itself, it is usually coloured by means of fillers and pigments. It was first used in sculpture in conjunction with powdered metal fillers in order to produce cheap "cold-cast" substitutes for bronze and aluminum, but with the recent tendency to use bright colours in sculpture it is now often coloured either by pigmenting the material itself or by painting.

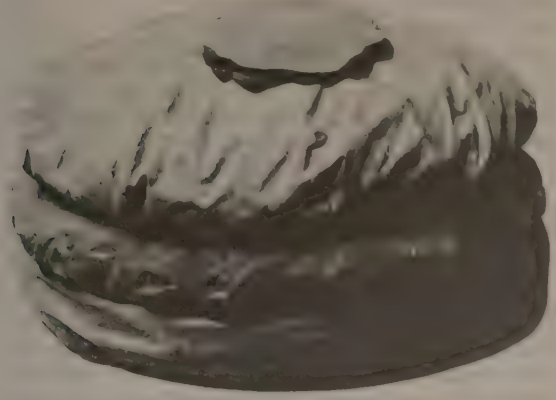
It is possible to model fibreglass, but more usually it is cast as a laminated shell. Its possibilities for sculpture have not yet been fully exploited.

Various formulas for modelling wax have been used in the past, but these have been generally replaced by synthetic waxes. The main uses of wax in sculpture have been as a preliminary modelling material for metal casting by the lost-wax, or cire perdue, process (see *Methods and techniques,* below) and for making sketches. It is not durable enough for use as a material in its own right, although it has been used for small works, such as wax fruit, that can be kept under a glass dome.

Papier-mâché (pulped paper bonded with glue) has been used for sculpture, especially in the Far East. Mainly used for decorative work, especially masks, it can have considerable strength; the Japanese, for example, made armour from it. Sculpture made of sheet paper is a limited art form used only for ephemeral and usually trivial work.

Numerous other permanent materials—such as shells, amber, and brick—and ephemeral ones—such as feathers, baker's dough, ice and snow, and cake icing—have been

used for fashioning three-dimensional images. In view of recent trends in sculpture it is no longer possible to speak of "the materials of sculpture." Modern sculpture has no special materials. Any material, natural or man-made, is likely to be used, including inflated polyethylene, foam rubber, expanded polystyrene, fabrics, and neon tubes; the materials for a mid-20th-century sculpture by Claes Oldenburg, for example, are listed as canvas, cloth, Dacron, metal, foam rubber, and Plexiglas. Real objects, too, may be incorporated in sculpture, as in the mixed-medium compositions of Edward Kienholz; even junk has its devotees, who fashion "junk" sculpture.



Unconventional materials of modern sculpture.
"Giant Hamburger," painted sailcloth stuffed with foam rubber, by Claes Oldenburg, 1962. In the Art Gallery of Ontario. 132 × 213 cm.

By courtesy of the Art Gallery of Ontario

Methods and techniques

Although a sculptor may specialize in, say, stone carving or direct metalwork, the art of sculpture is not identifiable with any particular craft or set of crafts. It presses into its service whatever crafts suit its purposes. Technologies developed for more utilitarian purposes are often easily adapted for sculpture; in fact, useful artifacts and sculptured images have often been produced in the same workshop, sometimes by the same craftsman. The methods and techniques employed in producing a pot, a bronze harness trapping, a decorative stone molding or column, a carved wooden newel post, or even a fibreglass car body are essentially the same as those used in sculpture. For example, the techniques of repoussé, metal casting, blacksmithing, sheet-metal work, and welding, which are used for the production of functional artifacts and decorative metalwork, are also used in metal sculpture; and the preparation, forming, glazing, decoration, and firing of clay are basically the same in both utilitarian pottery and pottery sculpture. The new techniques used by sculptors today are closely related to new techniques applied in building and industrial manufacture.

THE SCULPTOR AS DESIGNER AND AS CRAFTSMAN

The conception of an artifact or a work of art—its form, imaginative content, and expressiveness—is the concern of a designer, and it should be distinguished from the execution of the work in a particular technique and material, which is the task of a craftsman. A sculptor usually functions as both designer and craftsman, but these two aspects of sculpture may be separated.

Expressiveness of direct handling

Certain types of sculpture depend considerably for their aesthetic effect on the way in which their material has been directly manipulated by the artist himself. The direct, expressive handling of clay in a model by Rodin, or the use of the chisel in the stiacciato (very low) reliefs of the 15th-century Florentine sculptor Donatello could no more have been delegated to a craftsman than could the brushwork of Rembrandt. The actual physical process of working materials is for many sculptors an integral part of the art of sculpture, and their response to the working qualities of the material—such as its plasticity, hardness, and texture—is evident in the finished work. Design and craftsmanship are intimately fused in such a work, which is a highly personal expression.

Even when the direct handling of material is not as vital as this to the expressiveness of the work, it still may be impossible to separate the roles of the artist as designer and craftsman. The qualities and interrelationships of forms may be so subtle and complex that they cannot be adequately specified and communicated to a craftsman. Moreover, many aspects of the design may actually be contributed during the process of working. Michelangelo's way of working, for example, enabled him to change his mind about important aspects of composition as the work proceeded.

A complete fusion of design and craftsmanship may not be possible if a project is a large one or if the sculptor is too old or too weak to do all of the work himself. The sheer physical labour of making a large sculpture can be considerable, and sculptors from Phidias in the 5th century BC to Henry Moore in the 20th century have employed pupils and assistants to help with it. Usually the sculptor delegates the time-consuming first stages of the work or some of its less important parts to his assistants and executes the final stages or the most important parts himself.

On occasion, a sculptor may function like an architect or industrial designer. He may do no direct work at all on the finished sculpture, his contribution being to supply exhaustive specifications in the form of drawings and perhaps scale models for a work that is to be entirely fabricated by craftsmen. Obviously, such a procedure excludes the possibility of direct, personal expression through the handling of the materials; thus, works of this kind usually have the same anonymous, impersonal quality as architecture and industrial design. An impersonal approach to sculpture was favoured by many sculptors of the 1960s such as William Tucker, Donald Judd, and William Turnbull. They used the skilled anonymous workmanship of industrial fabrications to make their large-scale, extremely precise, simple sculptural forms that are called "primary structures."

Stages in the production of a major work

General methods. Broadly speaking, the stages in the production of a major work of sculpture conform to the following pattern: the commission; the preparation, submission, and acceptance of the design; the selection and preparation of materials; the forming of materials; surface finishing; installation or presentation.

Almost all of the sculpture of the past and some present-day sculpture originates in a demand made upon the sculptor from outside, usually in the form of a direct commission or through a competition. If the commission is for a portrait or a private sculpture, the client may only require to see examples of the artist's previous work; but if it is a public commission, the sculptor is usually expected to submit drawings and maquettes (small-scale, three-dimensional sketch models) that give an idea of the nature of the finished work and its relation to the site. He may be free to choose his own subject matter or theme, or it may be more or less strictly prescribed. A medieval master sculptor, for example, received the program for a complex scheme of church sculpture from theological advisers, and Renaissance contracts for sculpture were often extremely specified and detailed. Today a great deal of sculpture is not commissioned. It arises out of the sculptor's private concern with form and imagery, and he works primarily to satisfy himself. When the work is finished he may exhibit and attempt to sell it in an art gallery.

Most of the materials used by 20th-century sculptors are readily available in a usable form from builders' or sculptors' suppliers, but certain kinds of sculpture may involve a good deal of preparatory work on the materials. A sculptor may visit a stone quarry in order to select the material for a large project and to have it cut into blocks of the right size and shape. And since stone is costly to transport and best carved when freshly quarried, he may decide to do all of his work at the quarry. Because stone is extremely heavy, the sculptor must have the special equipment required for manoeuvring even small blocks into position for carving. A wood carver requires a supply of well-seasoned timber and may keep a quantity of logs

and blocks in store. A modeller needs a good supply of clay of the right kind. For large terra-cottas he may require a specially made-up clay body, or he may work at a brickworks, using the local clay and firing in the brick kilns.

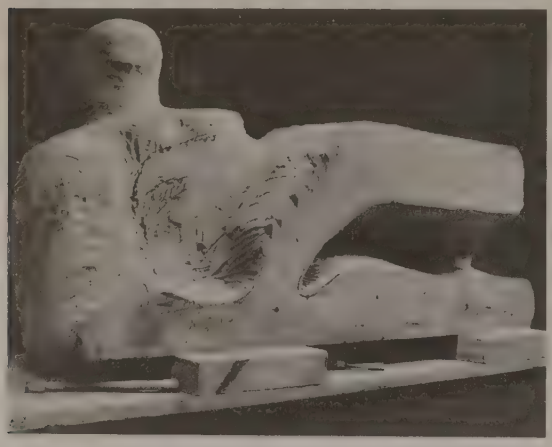
The main part of the sculptor's work, the shaping of the material itself by modelling, carving, or constructional techniques, may be a long and arduous process, perhaps extending over a number of years and requiring assistants. Much of the work, especially architectural decoration, may be carried out at the site, or in situ.

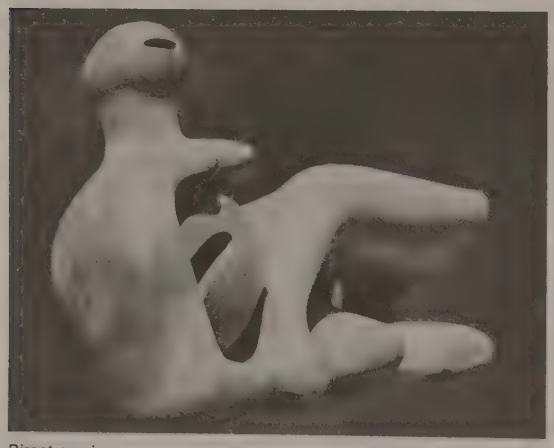
To improve its weathering qualities, to bring out the characteristics of its material to the best advantage, or to make it more decorative or realistic, sculpture is usually given a special surface finish. It may be rubbed down and polished, patinated, metal plated, gilded, painted, inlaid with other materials, and so on.

Installation of sculpture

Finally, the installation of sculpture may be a complex and important part of the work. The positioning and fixing of large architectural sculpture may involve cooperation with builders and engineers; fountains may involve elaborate plumbing; the design and placing of outdoor bases, or plinths, in relation to the site and the spectator may require careful thought. The choice of the materials, shape, and proportions of the base even for a small work requires a considerable amount of care.

Carving. Whatever material is used, the essential features of the direct method of carving are the same; the sculptor starts with a solid mass of material and reduces it systematically to the desired form. After he has blocked out the main masses and planes that define the outer limits of the forms, he works progressively over the whole sculpture, first carving the larger containing forms and planes and then the smaller ones until eventually the surface details are reached. Then he gives the surface whatever finish is required. Even with a preliminary model as a guide, the sculptor's concept constantly evolves and clarifies as the work proceeds; thus, as he adapts his design to the nature of the carving process and the material, his work develops as an organic whole.





Direct carving.
(Top) "Reclining Figure" in progress, elmwood sculpture by Henry Moore, 1945–46. (Bottom) "Reclining Figure." Length 1.90 m.

The process of direct carving imposes a characteristic order on the forms of sculpture. The faces of the original block, slab, or cylinder of material can usually still be sensed, existing around the finished work as a kind of implied spatial envelope limiting the extension of the forms in space and connecting their highest points across space. In a similar way, throughout the whole carving, smaller forms and planes can be seen as contained within implied larger ones. Thus, an ordered sequence of containing forms and planes, from the largest to the smallest, gives unity to the work.

Indirect carving. All of the great sculptural traditions of the past used the direct method of carving, but in Western civilization during the 19th and early 20th centuries it became customary for stone and, to a lesser extent, wood sculpture to be produced by the indirect method. This required the production of a finished clay model that was subsequently cast in plaster and then reproduced in stone or wood in a more or less mechanical way by means of a pointing machine (see *Reproduction and surface-finishing techniques: pointing* below). Usually, the carving was not done by the sculptor himself. At its worst, this procedure results in a carved copy of a design that was conceived in terms of clay modelling. Although indirect carving does not achieve aesthetic qualities that are typical of carved sculpture, it does not necessarily result in bad sculpture. Rodin's marble sculptures, for example, are generally considered great works of art even by those who object to the indirect methods by which they were produced. The indirect method has been steadily losing ground since the revival of direct carving in the early part of the 20th century, and today it is in general disrepute among carvers.

Tools for stone carving

Carving tools and techniques. The tools used for carving differ with the material to be carved. Stone is carved mostly with steel tools that resemble cold chisels. To knock off the corners and angles of a block, a tool called a pitcher is driven into the surface with a heavy iron hammer. The pitcher is a thick, chisel-like tool with a wide bevelled edge that breaks rather than cuts the stone. The heavy point then does the main roughing out, followed by the fine point, which may be used to within a short distance of the final surface. These pointed tools are hammered into the surface at an angle that causes the stone to break off in chips of varying sizes. Claw chisels, which have toothed edges, may then be worked in all directions over the surface, removing the stone in granule form and thus refining the surface forms. Flat chisels are used for finishing the surface carving and for cutting sharp detail. There are many other special tools, including stone gouges, drills, toothed hammers (known as bushhammers or bouchardes), and, often used today, power-driven pneumatic tools, for pounding away the surface of the stone.

Because medieval carvers worked mostly in softer stones and made great use of flat chisels, their work tends to have an edgy, cut quality and to be freely and deeply carved. In contrast is the work done in hard stones by people who lacked metal tools hard enough to cut the stone. Egyptian granite sculpture, for example, was produced mainly by abrasion; that is, by pounding the surface and rubbing it down with abrasive materials. The result is a compact sculpture, not deeply hollowed out, with softened edges and flowing surfaces. It usually has a high degree of tactile appeal.

Although the process of carving is fundamentally the same for wood or stone, the physical structure of wood demands tools of a different type. For the first blocking out of a wood carving a sculptor may use saws and axes, but his principal tools are a wide range of wood-carver's gouges. The sharp, curved edge of a gouge cuts easily through the bundles of fibre and when used properly will not split the wood. Flat chisels are also used, especially for carving sharp details. Wood rasps, or coarse files, and sandpaper can be used to give the surface a smooth finish, or, if preferred, it can be left with a faceted, chiselled appearance. Wood-carving tools have hardwood handles and are struck with round, wooden mallets. African wood

sculptors use a variety of adzes rather than gouges and mallets. Ivory is carved with an assortment of saws, knives, rasps, files, chisels, drills, and scrapers.

Modelling. In contrast to the reductive process of carving, modelling is essentially a building-up process in which the sculpture grows organically from the inside. Numerous plastic materials are used for modelling. The main ones are clay, plaster, and wax; but concrete, synthetic resins, plastic wood, stucco, and even molten metal can also be modelled. A design modelled in plastic materials may be intended for reproduction by casting in more permanent and rigid materials, such as metal, plaster, concrete, and fibreglass, or it may itself be made rigid and more permanent through the self-setting properties of its materials (for example, plaster) or by firing.

Clay models

Modelling for casting. The material most widely used for making positive models for casting is clay. A small, compact design or a low relief can be modelled solidly in clay without any internal support; but a large clay model must be formed over a strong armature made of wood and metal. Since the armature may be very elaborate and can only be altered slightly, if at all, once work has started, the modeller must have a fairly clear idea from his drawings and maquettes of the arrangement of the main shapes of the finished model. The underlying main masses of the sculpture are built up firmly over the armature, and then the smaller forms, surface modelling, and details are modelled over them. The modeller's chief tools are his fingers, but for fine work he may use a variety of wooden modelling tools to apply the clay and wire loop tools to cut it away. Reliefs are modelled on a vertical or nearly vertical board. The clay is keyed, or secured, onto the board with galvanized nails or wood laths. The amount of armature required depends on the height of the relief and the weight of clay involved.

To make a cast in metal, a foundry requires from the sculptor a model made of a rigid material, usually plaster. The sculptor can produce this either by modelling in clay and then casting in plaster from the clay model or by modelling directly in plaster. For direct plaster modelling, a strong armature is required because the material is brittle. The main forms may be built up roughly over the armature in expanded wire and then covered in plaster-soaked scrim (a loosely woven sacking). This provides a hollow base for the final modelling, which is done by applying plaster with metal spatulas and by scraping and cutting down with rasps and chisels.

Fibreglass and concrete sculptures are cast in plaster molds taken from the sculptor's original model. The model is usually clay rather than plaster because if the forms of the sculpture are at all complex it is easier to remove a plaster mold from a soft clay model than from a model in a rigid material, such as plaster.

A great deal of the metal sculpture of the past, including Nigerian, Indian, and many Renaissance bronzes, was produced by the direct lost-wax process, which involves a special modelling technique (see *Reproduction and surface-finishing techniques: casting and molding* below). The design is first modelled in some refractory material to within a fraction of an inch of the final surface, and then the final modelling is done in a layer of wax, using the fingers and also metal tools, which can be heated to make the wax more pliable. Medallions are often produced from wax originals, but because of their small size they do not require a core.

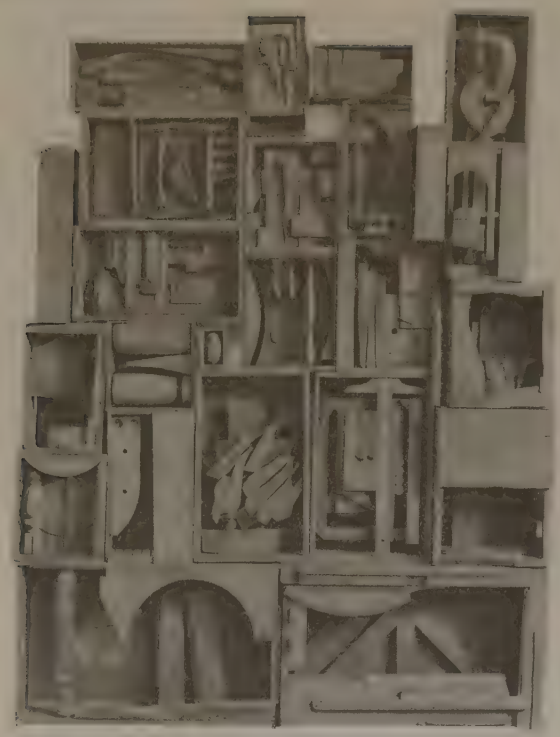
Modelling for pottery sculpture. To withstand the stresses of firing, a large pottery sculpture must be hollow and of an even thickness. There are two main ways of achieving this. In the process of hollow modelling, which is typical of the potter's approach to form, the main forms of the clay model are built up directly as hollow forms with walls of a roughly even thickness. The methods of building are similar to those employed for making hand-built pottery—coiling, pinching, and slabbing. The smaller forms and details are then added, and the finished work is allowed to dry out slowly and thoroughly before firing. The process of solid modelling is more typical of the sculptor's traditional approach to form. The sculpture is modelled in solid clay, sometimes over a carefully considered armature, by the sculptor's usual methods of clay modelling. Then it is cut open and hollowed out, and the armature, if there is one, is removed. The pieces are then rejoined and the work is dried out and fired.

Effects of modelling on the design of sculpture

General characteristics of modelled sculpture. The process of modelling affects the design of sculpture in three important ways. First, the forms of the sculpture tend to be ordered from the inside. There are no external containing forms and planes, as in carved sculpture. The overall design of the work—its main volumes, proportions, and axial arrangement—is determined by the underlying forms; and the smaller forms, surface modelling, and decorative details are all formed around and sustained by this underlying structure. Second, because its extension into space is not limited by the dimensions of a block of material, modelled sculpture tends to be much freer and more expansive in its spatial design than carved sculpture. If the tensile strength of metal is to be exploited in the finished work, there is almost unlimited freedom; designs for brittle materials such as concrete or plaster are more limited. Third, the plasticity of clay and wax encourages a fluent, immediate kind of manipulation, and many sculptors, such as Auguste Rodin, Giacomo Manzù, and Sir Jacob Epstein, like to preserve this record of their direct handling of the medium in their finished work. Their approach contrasts with that of the Benin and Indian bronze sculptors, who refined the surfaces of their work to remove all traces of personal "handwriting."

Lost-wax process bronze sculpture.
"Male Portrait Head," bronze from Ife, Nigeria, 12th century. In the Museum of Ife Antiquities. Height 34 cm.
Eliot Elisofon

Constructing and assembling. A constructed or assembled sculpture is made by joining preformed pieces of material. It differs radically in principle from carved and modelled sculpture, both of which are fabricated out of a homogeneous mass of material. Constructed sculpture is made out of such basic preformed components as metal tubes, rods, plates, bars, and sheets; wooden laths, planks, dowels, and blocks; laminated timbers and chipboards; sheets of Perspex, Formica, and glass; fabrics; and wires and threads. These are cut to various sizes and may be either shaped before they are assembled or used as they are. The term assemblage is usually reserved for constructed sculpture that incorporates any of a vast array of ready-made, so-called found objects, such as old boilers, typewriters, engine components, mirrors, chairs, and table legs and other bits of old



Assemblage.
"Black Wall," wood sculpture by Louise Nevelson, 1959. In the Tate Gallery, London. 2.84 × 2.16 × 0.65 m.
By courtesy of the trustees of the Tate Gallery, London

furniture. Numerous techniques are employed for joining these components, most of them derived from crafts other than traditional sculptural ones; for example, metal welding and brazing, wood joinery, bolting, screwing, rivetting, nailing, and bonding with new powerful adhesives.

The use of constructional techniques to produce sculpture is the main technical development of the art in recent years. Among the reasons for its popularity are that it lends itself readily to an emphasis on the spatial aspects of sculpture that have preoccupied so many 20th-century artists; it is quicker than carving and modelling; it is considered by many sculptors and critics to be especially appropriate to a technological civilization; it is opening up new fields of imagery and new types of symbolism and form.

For constructed "gallery" sculpture, almost any materials and techniques are likely to be used, and the products are often extremely ephemeral. But architectural sculpture, outdoor sculpture, and indeed any sculpture that is actually used must be constructed in a safe and at least reasonably permanent manner. The materials and techniques employed are therefore somewhat restricted. Metal sculpture constructed by rivetting, bolting, and, above all, welding and brazing is best for outdoor use.

Direct metal sculpture. The introduction of the oxyacetylene welding torch as a sculptor's tool has revolutionized metal sculpture in recent years. A combination of welding and forging techniques was pioneered by the Spanish sculptor Julio González around 1930; and during the 1940s and 1950s it became a major sculptural technique, particularly in Britain and in the United States, where its greatest exponent was David Smith. In the 1960s and early 1970s, more sophisticated electric welding processes were replacing flame welding.

Welding and forging

Welding equipment can be used for joining and cutting metal. A welded joint is made by melting and fusing together the surfaces of two pieces of metal, usually with the addition of a small quantity of the same metal as a filler. The metal most widely used for welded sculpture is mild steel, but other metals can be welded. In a brazed joint, the parent metals are not actually fused together but are joined by an alloy that melts at a lower temperature than the parent metals. Brazing is particularly useful for making joints between different kinds of metal, which cannot be done by welding, and for joining nonferrous metals. Forging is the direct shaping of metal by bending, hammering, and cutting.

Direct metalworking techniques have opened up whole new ranges of form to the sculptor—open skeletal structures, linear and highly extended forms, and complex, curved sheet forms. Constructed metal sculpture may be precise and clean, as that of Minimalist sculptors Donald Judd and Phillip King, or it may exploit the textural effects of molten metal in a free, "romantic" manner.

REPRODUCTION AND SURFACE-FINISHING TECHNIQUES

Casting and molding processes are used in sculpture either for making copies of existing sculpture or as essential stages in the production of a finished work. Numerous materials are used for making molds and casts, and some of the methods are complex and highly skilled. Only a broad outline of the principal methods can be given here.

Casting and molding. These are used for producing a single cast from a soft, plastic original, usually clay. They are especially useful for producing master casts for subsequent reproduction in metal. The basic procedure is as follows. First, the mold is built up in liquid plaster over the original clay model; for casting reliefs, a one-piece mold may be sufficient, but for sculpture in the round a mold in at least two sections is required. Second, when the plaster is set, the mold is divided and removed from the clay model. Third, the mold is cleaned, reassembled, and filled with a self-setting material such as plaster, concrete, or fibreglass-reinforced resin. Fourth, the mold is carefully chipped away from the cast. This involves the destruction of the mold—hence the term "waste" mold. The order of reassembling and filling the mold may be reversed; fibreglass and resin, for example, are "laid up" in the mold pieces before they are reassembled.

Plaster piece molds are used for producing more than one cast from a soft or rigid original and are especially good for reproducing existing sculpture and for slip casting (see below). Before the invention of flexible molds (see below), piece molds were used for producing wax casts for metal casting by the lost-wax process. A piece mold is built up in sections that can be withdrawn from the original model without damaging it. The number of sections depends on the complexity of the form and on the amount of undercutting; tens, or even hundreds, of pieces may be required for really large, complex works. The mold sections are carefully keyed together and supported by a plaster case. When the mold has been filled, it can be removed section by section from the cast and used again. Piece molding is a highly skilled and laborious process.

Made of such materials as gelatin, vinyl, and rubber, flexible molds are used for producing more than one cast; they offer a much simpler alternative to piece molding when the original model is a rigid one with complex forms and undercuts. The material is melted and poured around the original positive in sections, if necessary. Being flexible, the mold easily pulls away from a rigid surface without causing damage. While it is being filled (with wax, plaster, concrete, and fibreglass-reinforced resins), the mold must be surrounded by a plaster case to prevent distortion.

The traditional method of casting metal sculpture

The lost-wax process is the traditional method of casting metal sculpture. It requires a positive, which consists of a core made of a refractory material and an outer layer of wax. The positive can be produced either by direct modelling in wax over a prepared core, in which case the process is known as direct lost-wax casting, or by casting in a piece mold or flexible mold taken from a master cast. The wax positive is invested with a mold made of refractory materials and is then heated to a temperature that will drive off all moisture and melt out all the wax, leaving a narrow cavity between the core and the investment. Molten metal is then poured into this cavity. When the metal has cooled down and solidified, the investment is broken away, and the core is removed from inside the cast. The process is, of course, much more complex than this simple outline suggests. Care has to be taken to suspend the core within the mold by means of metal pins, and a structure of channels must be made in the mold that will enable the metal to reach all parts of the cavity and

permit the mold gases to escape. A considerable amount of filing and chasing of the cast is usually required after casting is completed.

While the lost-wax process is used for producing complex, refined metal castings, sand molding is more suitable for simpler types of form and for sculpture in which a certain roughness of surface does not matter. Recent improvements in the quality of sand castings and the invention of the "lost-pattern" process (see below) have resulted in a much wider use of sand casting as a means of producing sculpture. A sand mold, made of special sand held together by a binder, is built up around a rigid positive, usually in a number of sections held together in metal boxes. For a hollow casting, a core is required that will fit inside the negative mold, leaving a narrow cavity as in the lost-wax process. The molten metal is poured into this cavity.

The lost-pattern process is used for the production by sand molding of single casts in metal. After a positive made of expanded polystyrene is firmly embedded in casting sand, molten metal is poured into the mold straight onto the expanded foam original. The heat of the metal causes the foam to pass off into vapour and disappear, leaving a negative mold to be filled by the metal. Channels for the metal to run in and for the gases to escape are made in the mold, as in the lost-wax process. The method is used mainly for producing solid castings in aluminum that can be welded or rivetted together to make the finished sculpture.

Slip casting is primarily a potter's technique that can be used for repetition casting of small pottery sculptures. Liquid clay, or slip, is poured into a plaster piece mold. Some of the water in the slip is absorbed by the plaster and a layer of stiffened clay collects on the surface of the mold. When this layer is thick enough to form a cast, the excess slip is poured off and the mold is removed. The hollow clay cast is then dried and fired.

Simple casts for pottery sculpture—mainly tiles and low reliefs—can be prepared by pressing clay into a rigid mold. More complex forms can be built up from a number of separately press-cast pieces. Simple terra-cotta molds can be made by pressing clay around a rigid positive form. After firing, these press molds can be used for press casting.

Pointing. A sculpture can be reproduced by transposing measurements taken all over its surface to a copy. The process is made accurate and thorough by the use of a pointing machine, which is an arrangement of adjustable metal arms and pointers that are set to the position of any point on the surface of a three-dimensional form and then used to locate the corresponding point on the surface of a copy. If the copy is a stone one, the block is drilled to the depth measured by the pointing machine. When a number of points have been fixed by drilling, the stone is cut away to the required depth. For accurate pointing, a vast number of points have to be taken, and the final surface is approached gradually. The main use of pointing has been for the indirect method of carving.

Enlarged and reduced copies

Enlarged and reduced copies of sculpture can also be produced with the aid of mechanical devices. A sophisticated reducing machine that works on the principle of the pantograph (an instrument for copying on any predetermined scale, consisting of four light, rigid bars jointed in parallelogram form) is used in minting for scaling down the sculptor's original model to coin size.

Surface finishing. Surface finishes for sculpture can be either natural—bringing the material of the sculpture itself to a finish—or applied. Almost all applied surface finishes preserve as well as decorate.

Smoothing and polishing. Many sculptural materials have a natural beauty of colour and texture that can be brought out by smoothing and polishing. Stone carvings are smoothed by rubbing down with a graded series of coarse and fine abrasives, such as carborundum, sandstone, emery, pumice, and whiting, all used while the stone is wet. Some stones, such as marble and granite, will take a high gloss; others are too coarse-grained to be polished and can only be smoothed to a granular finish. Wax is sometimes used to give stone a final polish.

The natural beauty of wood is brought out by sandpapering or scraping and then waxing or oiling. Beeswax and linseed oil are the traditional materials, but a wide range of waxes and oils is currently available.

Ivory is polished with gentle abrasives such as pumice and whiting, applied with a damp cloth.

Concrete can be rubbed down, like stone, with water and abrasives, which both smooth the surface and expose the aggregate. Some concretes can be polished.

Metals are rubbed down manually with steel wool and emery paper and polished with various metal polishes. A high-gloss polish can be given to metals by means of power-driven buffing wheels used in conjunction with abrasives and polishes. Clear lacquers are applied to preserve the polish.

Painting. Stone, wood, terra-cotta, metal, fibreglass, and plaster can all be painted in a reasonably durable manner provided that the surfaces are properly prepared and suitable primings and paints are used. In the past, stone and wood carvings were often finished with a coating of gesso (plaster of Paris or gypsum prepared with glue) that served both as a final modelling material for delicate surface detail and as a priming for painting. Historically, the painting and gilding of sculpture were usually left to specialists. In Greek relief sculpture, actual details of the composition were often omitted at the carving stage and left for the painter to insert. In the 15th century, the great Flemish painter Rogier van der Weyden undertook the painting of sculpture as part of his work.

Modern paint technology has made an enormous range of materials available. Constructed sculptures are often finished with mechanical grinders and sanders and then sprayed with high-quality cellulose paints.

Gilding. The surfaces of wood, stone, and plaster sculpture can be decorated with gold, silver, and other metals that are applied in leaf or powder form over a suitable priming. Metals, especially bronze, were often fire-gilded; that is, treated with an amalgam of gold and mercury that was heated to drive off the mercury. The panels of the "Gates of Paradise" in Florence, by the 15th-century sculptor Lorenzo Ghiberti, are a well-known example of gilded bronze.

Patination. Patinas on metals are caused by the corrosive action of chemicals. Sculpture that is exposed to different kinds of atmosphere or buried in soil or immersed in seawater for some time acquires a patina that can be extremely attractive. Similar effects can be achieved artificially by applying various chemicals to the metal surface. This is a particularly effective treatment for bronze, which can be given a wide variety of attractive green, brown, blue, and black patinas. Iron is sometimes allowed to rust until it acquires a satisfactory colour, and then the process is arrested by lacquering.

Patinas achieved artificially

Electroplating. The surfaces of metal sculpture or of specially prepared nonmetal sculpture can be coated with such metals as chrome, silver, gold, copper, and nickel by the familiar industrial process of electroplating. The related technique of anodizing can be used to prevent the corrosion of aluminum sculpture and to dye its surface.

Other finishes. The surfaces of metal sculpture can be decorated by means of numerous metalsmithing techniques—etching, engraving, metal inlaying, enamelling, and so on. Pottery sculpture can be decorated with coloured slips, oxides, and enamels; glazed with a variety of shiny or mat glazes; and brought to a dull polish by burnishing.

Other materials have often been added to the surface of sculpture. The eyes of ancient figure sculpture, for example, were sometimes inlaid with stones. Occasionally—as in Mexican mosaic work—the whole surface of a sculpture is inlaid with mother-of-pearl, turquoise, coral, and many other substances.

Forms, subject matter, imagery, and symbolism of sculpture

A great deal of sculpture is designed to be placed in public squares, gardens, parks, and similar open places or in interior positions where it is isolated in space and can be viewed from all directions. Other sculpture is carved in relief and is viewed only from the front and sides.

SCULPTURE IN THE ROUND

The opportunities for free spatial design that such freestanding sculpture presents are not always fully exploited. The work may be designed, like many Archaic sculptures, to be viewed from only one or two fixed positions, or it may in effect be little more than a four-sided relief that hardly changes the three-dimensional form of the block at all. Sixteenth-century Mannerist sculptors, on the other hand, made a special point of exploiting the all-around visibility of freestanding sculpture. Giambologna's "Rape of the Sabines," for example, compels the viewer to walk all around it in order to grasp its spatial design. It has no principal views; its forms move around the central axis of the composition, and their serpentine movement unfolds itself gradually as the spectator moves around to follow them. Much of the sculpture of Henry Moore and other 20th-century sculptors is not concerned with movement of this kind, nor is it designed to be viewed from any fixed positions. Rather, it is a freely designed structure of multidirectional forms that is opened up, pierced, and extended in space in such a way that the viewer is made aware of its all-around design largely by seeing through the sculpture. The majority of constructed sculptures are disposed in space with complete freedom and invite viewing from all directions. In many instances the spectator can actually walk under and through them.

The way in which a freestanding sculpture makes contact with the ground or with its base is a matter of considerable importance. A reclining figure, for example, may in effect be a horizontal relief. It may blend with the ground plane and appear to be rooted in the ground like an outcrop of rock. Other sculptures, including some reclining figures, may be designed in such a way that they seem to rest on the ground and to be independent of their base. Others are supported in space above the ground. The most completely freestanding sculptures are those that have no base and may be picked up, turned in the hands, and literally viewed all around like a netsuke (a small toggle of wood, ivory, or metal used to fasten a small pouch or purse to a kimono sash). Of course, a large sculpture cannot actually be picked up in this way, but it can be designed so as to invite the viewer to think of it as a detached, independent object that has no fixed base and is designed all around.

Alinari—Art Resource

Sculpture in the round.
The changing appearance of freestanding sculpture observed when walking around a statue is suggested by these two views of "Rape of the Sabines," marble sculpture by Giambologna, 1583. In the Loggia dei Lanzi, Florence. Height 411 cm.

Sculpture designed to stand against a wall or similar background or in a niche may be in the round and freestanding in the sense that it is not attached to its background like a relief; but it does not have the spatial independence of completely freestanding sculpture, and it is not designed to be viewed all around. It must be designed so that its formal structure and the nature and meaning of its subject matter can be clearly apprehended from a limited range of frontal views. The forms of the sculpture, therefore, are usually spread out mainly in a lateral direction rather than in depth. Greek pedimental sculpture illustrates this approach superbly: the composition is spread out in a plane perpendicular to the viewer's line of sight and is made completely intelligible from the front. Seventeenth-century Baroque sculptors, especially Bernini, adopted a rather different approach. There may be considerable recession and foreshortening in their compositions, but the forms are carefully arranged so that they present a coherent and intelligible whole from one special frontal viewpoint.

Difference between wall and niche and freestanding sculpture

The frontal composition of wall and niche sculpture does not necessarily imply any lack of three-dimensionality in the forms themselves; it is only the arrangement of the forms that is limited. Classical pedimental sculpture, Indian temple sculpture such as that at Khajurāho, Gothic niche sculpture, and Michelangelo's Medici tomb figures are all designed to be placed against a background, but their forms are conceived with a complete fullness of volume.

RELIEF SCULPTURE

Relief sculpture is a complex art form that combines many features of the two-dimensional pictorial arts and the three-dimensional sculptural arts. On the one hand, a relief, like a picture, is dependent on a supporting surface, and its composition must be extended in a plane in order to be visible. On the other hand, its three-dimensional properties are not merely represented pictorially but are in some degree actual, like those of fully developed sculpture.

Among the various types of relief are some that approach very closely the condition of the pictorial arts. The reliefs of Donatello, Ghiberti, and other early Renaissance artists make full use of perspective, which is a pictorial method of representing three-dimensional spatial relationships realistically on a two-dimensional surface. Egyptian and most pre-Columbian American low reliefs are also extremely pictorial but in a different way. Using a system of graphic conventions, they translate the three-dimensional world into a two-dimensional one. The relief image is essentially one of plane surfaces and could not possibly exist in three dimensions. Its only sculptural aspects are its slight degree of actual projection from a surface and its frequently subtle surface modelling.

Other types of relief—for example, Classical Greek and most Indian—are conceived primarily in sculptural terms. The figures inhabit a space that is defined by the solid forms of the figures themselves and is limited by the background plane. This back plane is treated as a finite, impenetrable barrier in front of which the figures exist. It is not conceived as a receding perspective space or environment within which the figures are placed nor as a flat surface upon which they are placed. The reliefs, so to speak, are more like contracted sculpture than expanded pictures.

The central problem of relief sculpture is to contract or condense three-dimensional solid form and spatial relations into a limited depth space. The extent to which the forms actually project varies considerably, and reliefs are classified on this basis as low reliefs (bas-reliefs) or high reliefs. There are types of reliefs that form a continuous series from the almost completely pictorial to the almost fully in the round.

One of the relief sculptor's most difficult tasks is to

Representation of forms in depth

represent the relations between forms in depth within the limited space available to him. He does this mainly by giving careful attention to the planes of the relief. In a carved relief the highest, or front, plane is defined by the surface of the slab of wood or stone in which the relief is carved; and the back plane is the surface from which the forms project. The space between these two planes can be thought of as divided into a series of planes, one behind the other. The relations of forms in depth can then be thought of as relations between forms lying in different planes.

Sunken relief is also known as incised, coelanaglyphic, and intaglio relief. It is almost exclusively an ancient Egyptian art form, but some beautiful small-scale Indian examples in ivory have been discovered at Bagrām in Afghanistan. In a sunken relief, the outline of the design is first incised all around. The relief is then carved inside the incised outline, leaving the surrounding surface untouched. Thus, the finished relief is sunk below the level of the surrounding surface and is contained within a sharp, vertical-walled contour line. This approach to relief sculpture preserves the continuity of the material's original surface and creates no projection from it. The outline shows up as a powerful line of light and shade around the whole design.

Figurative low relief is generally regarded by sculptors as an extremely difficult art form. To give a convincing impression of three-dimensional structure and surface modelling with only a minimal degree of projection demands a fusion of draftsmanship and carving or modelling skill of a high order. The sculptor has to proceed empirically, constantly changing the direction of his light and testing the optical effect of his work. He cannot follow any fixed rules or represent things in depth by simply scaling down measurements mathematically, so that, say, one inch of relief space represents one foot.

The forms of low relief usually make contact with the background all around their contours. If there is a slight amount of undercutting, its purpose is to give emphasis, by means of cast shadow, to a contour rather than to give any impression that the forms are independent of their background. Low relief includes figures that project up to about half their natural circumference.

Technically, the simplest kind of low relief is the two-plane relief. For this, the sculptor draws an outline on a surface and then cuts away the surrounding surface, leaving the figure raised as a flat silhouette above the background plane. This procedure is often used for the first stages of a full relief carving, in which case the sculptor will proceed to carve into the raised silhouette, rounding the forms and giving an impression of three-dimensional structure. In a two-plane relief, however, the silhouette is left flat and substantially unaltered except for the addition of surface detail. Pre-Columbian sculptors used this method of relief carving to create bold figurative and abstract reliefs.

Stiacciato relief

Stiacciato relief is an extremely subtle type of flat, low relief carving that is especially associated with the 15th-century sculptors Donatello and Desiderio da Settignano. The design is partly drawn with finely engraved chisel lines and partly carved in relief. The stiacciato technique depends largely for its effect on the way in which pale materials, such as white marble, respond to light and show up the most delicate lines and subtle changes of texture or relief.

The forms of high relief project far enough to be in some degree independent of their background. As they approach the fullness of sculpture in the round, they become of necessity considerably undercut. In many high reliefs, where parts of the composition are completely detached from their background and fully in the round, it is often impossible to tell from the front whether or not a figure is actually attached to its background.

Many different degrees of projection are often combined in one relief composition. Figures in the foreground may be completely detached and fully in the round, while those in the middle distance are in about half relief and those in the background in low relief. Such effects are common in late Gothic, Renaissance, and Baroque sculpture.

By courtesy of the trustees of the British Museum

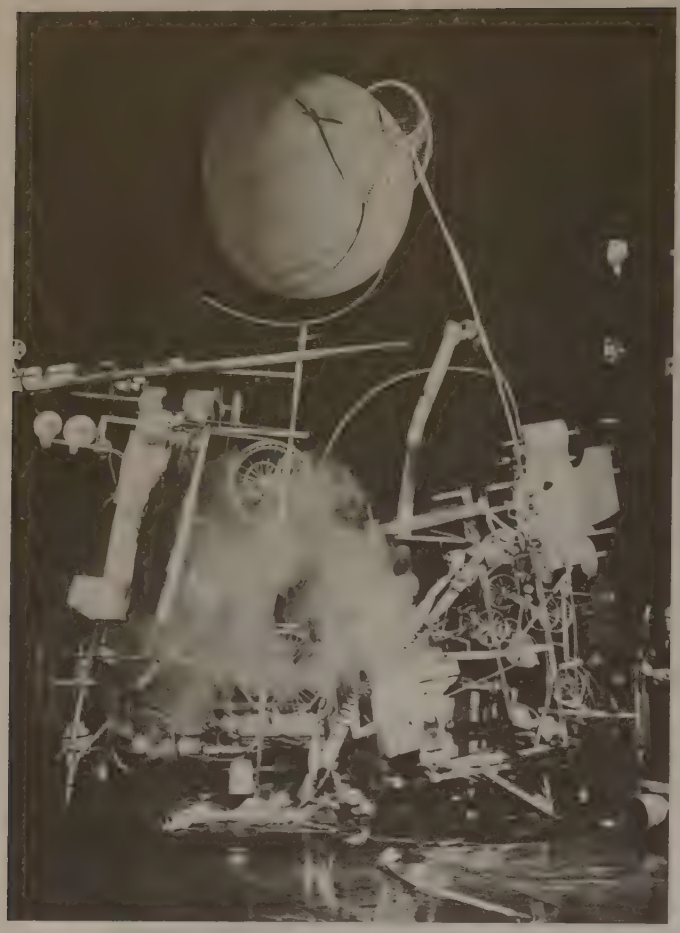
Sunken relief.
Egyptian woman carrying food for the dead, detail of an Old Kingdom limestone relief originally from the tomb of Thetha at Thebes, *c.* 2300 BC. In the British Museum. Figure height 43 cm.

MODERN FORMS OF SCULPTURE

Since the 1950s, many new combined forms of art have been developed that do not fit readily into any of the traditional categories. Two of the most important of these, environments and kinetics, are closely enough connected with sculpture to be regarded by many artists and critics as branches or offshoots of sculpture. It is likely, however, that the persistence of the terms environmental sculpture and kinetic sculpture is a result of the failure of language to keep pace with events; for the practice is already growing of referring simply to environments and kinetics, as one might refer to painting, sculpture, and engraving, as art forms in their own right.

Traditional sculptures in relief and in the round are static, fixed objects or images. Their immobility and immutability are part of the permanence traditionally associated with the art of sculpture, especially monumental sculpture. What one refers to as movement in, say, a Baroque or Greek sculpture is not actual physical motion but a movement that is either directly represented in the subject matter (galloping horses) or expressed through the dynamic character of its form (spirals, undulating curves). In recent years, however, the use of actual movement, kineticism, has become an important aspect of sculpture. Naum Gabo, Marcel Duchamp, László Moholy-Nagy, and Alexander Calder were pioneers of kinetic sculpture in modern times, but many kinetic artists see a connection between their work and such forms as the moving toys, dolls, and clocks of previous ages.

There are now types of sculpture in which the components are moved by air currents, as in the well-known mobiles of Calder; by water; by magnetism, the speciality of Nicholus Takis; by a variety of electromechanical devices; or by the participation of the spectator himself. The neo-



Kinetic sculpture.
"Homage to New York," a self-constructing and self-destroying work of art by Jean Tinguely, 1960. Small pieces now in the Museum of Modern Art, New York.
By courtesy of the Museum of Modern Art, New York; photograph, David Gahr

Dada satire quality of the kinetic sculpture created during the 1960s is exemplified by the works of Jean Tinguely. His self-destructing "Homage to New York" perfected the concept of a sculpture being both an object and an event, or "happening."

The aim of most kinetic sculptors is to make movement itself an integral part of the design of the sculpture and not merely to impart movement to an already complete static object. Calder's mobiles, for example, depend for their aesthetic effect on constantly changing patterns of relationship. When liquids and gases are used as components, the shapes and dimensions of the sculpture may undergo continual transformations. The movement of smoke; the diffusion and flow of coloured water, mercury, oil, and so on; pneumatic inflation and deflation; and the movement of masses of bubbles have all served as media for kinetic sculpture. In the complex, electronically controlled "spatio-dynamic" and "lumino-dynamic" constructions of Nicolas Schöffer, the projection of changing patterns of light into space is a major feature.

The environmental sculptor creates new spatial contexts that differ from anything developed by traditional sculpture. The work no longer confronts the spectator as an object but surrounds him so that he moves within it as he might within a stage set, a garden, or an interior. The most common type of environment is the "room," which may have specially shaped and surfaced walls, special lighting effects, and many different kinds of contents. Kurt Schwitters' *Merzbau* (destroyed in 1943) was the first of these rooms, which now include the nightmare fantasy of Edward Kienholz' tableaux, such as "Roxy's" (1961) or "The Illegal Operation" (1962); George Segal's compositions, in which casts of clothed human figures in frozen, casual attitudes are placed in interiors; and rooms built of mirrors, such as Yayoi Kusama's "Endless Love Room" and Lucas Samaras' "Mirrored Room," in both of which the spectator himself, endlessly reflected, becomes part of the total effect.

Environmental art, in common with collage and assemblage, has tended toward greater concreteness not by making a more realistic representation, as naturalistic art does, but by including more of reality itself in the work; for example, by using casts taken from the actual human body, real clothes, actual objects and casts of objects, actual lighting effects, and real items of furniture. Plastic elements may be combined with music and sound effects, dance, theatrical spectacles, and film to create so-called happenings, in which real figures are constituents of the "artwork" and operations are performed not on "artistic" materials but are performed on real objects and on the actual environment. Ideas such as these go far beyond anything that has ever before been associated with the term sculpture.

Use of real objects in environmental art

REPRESENTATIONAL SCULPTURE

Sculpture in the round is much more restricted than relief in the range of its subject matter. The representation of, say, a battle scene or a cavalcade in the round would require a space that corresponded in scale in every direction with that occupied by an actual battle or cavalcade. No such problems arise in relief because the treatment of scale and relations in depth is to some extent notional, or theoretical, like that of pictures. Then again, because a relief is attached to a background, problems of weight and physical balance and support do not arise. Figures can be represented as floating in space and can be arranged vertically as well as horizontally. Thus, in general, sculpture in the round is concerned with single figures and limited groups, while reliefs deal with more complex "pictorial" subjects involving crowds, landscape, architectural backgrounds, and so on.

The human figure. The principal subject of sculpture has always been the human figure. Next in importance are animals and fantastic creatures based on human and animal forms. Other subjects—for example, landscape, plants, still life, and architecture—have served primarily as accessories to figure sculpture, not as subjects in their own right. The overwhelming predominance of the human figure is due: first, to its immense emotional importance as an object of desire, love, fear, respect, and, in the case of anthropomorphic gods, worship; and, second, to its inexhaustible subtlety and variety of form and expression. The nude or almost nude figure played a prominent role in Egyptian, Indian, Greek, and African sculpture, while in medieval European and ancient Chinese sculpture the figure is almost invariably clothed. The interplay of the linear and modelled forms of free draperies with the solid volumes of the human body was of great interest to Classical sculptors and later became one of the principal themes of Renaissance and post-Renaissance sculpture. The human figure continues to be of central importance in modern sculpture in spite of the growth of nonfigurative art; but the optimistic, idealized, or naturalistic images of man prevalent in previous ages have been largely replaced by images of despair, horror, deformation, and satire.

Devotional images and narrative sculpture. The production of devotional images has been one of the sculptor's main tasks, and many of the world's greatest sculptures are of this kind. They include images of Buddha and the Hindu gods; of Christ, the Virgin, and the Christian saints; of Athena, Aphrodite, Zeus, and other Greek gods; and of all the various gods, spirits, and mythical beings of Rome, the ancient Near East, pre-Columbian America, Black Africa, and the Pacific Islands.

Closely connected with devotional images are all of the narrative sculptures in which legends, heroic deeds, and religious stories are depicted for the delight and instruction of peoples who lived when books and literacy were rare. The Buddhist, Hindu, and Christian traditions are especially rich in narrative sculpture. Stories of the incarnations of Buddha—*Jātaka*—and of the Hindu gods abounded in the temple sculpture of India and Southeast Asia; for example, at Sānchi, Amarāvatī, Borobuḍur, and

Environmental sculpture.
"Mirrored Room," mirror on wood by Lucas Samaras, 1966. In the Albright-Knox Art Gallery, Buffalo. 305 × 244 cm.

Albright-Knox Art Gallery, Buffalo, gift of Seymour H. Knox; photograph, by courtesy of the Pace Gallery, New York

Angkor. Sculpture illustrating the stories of the Bible is so abundant in medieval churches that the churches have been called "Bibles in stone." Sculpture recounting the heroic deeds of kings and generals are common, especially in Assyria and Rome. The Romans made use of a form known as continuous narrative, the best known example of which is the spiral, or helical, band of relief sculpture that surrounds Trajan's Column (*c.* AD 106–113) and tells the story of the Emperor's Dacian Wars. The episodes in the narrative are not separated into a series of framed compositions but are linked to form a continuous band of unbroken relief.

Portraiture. Portraiture was practiced by the Egyptians but was comparatively rare in the ancient world until the Romans made portrait sculpture one of their major artistic achievements. The features of many famous people are known to modern man only through the work of Roman sculptors on coins and medals, portrait busts, and full-length portraits. Portraiture has been an important aspect of Western sculpture from the Renaissance to the present day. Some of the best known modern portrait sculptors are Rodin, Charles Despiau, Marino Marini, and Jacob Epstein.

Modern portrait sculptors

Scenes of everyday life. Scenes of everyday life have been represented in sculpture mainly on a small scale in minor works. The sculptures that are closest in spirit to the quiet dignity of the great 17th- and 18th-century genre paintings of Jan Vermeer and Jean-Baptiste-Siméon Chardin are, perhaps certain Greek tombstones, such as that of the Stele of Hegeso, which represents a quiet, absorbed moment when a seated young woman and her maidservant are looking at a necklace they have just removed from a casket. Intimate scenes of the people and their activities in everyday rural life are often portrayed in medieval and Egyptian reliefs as part of larger compositions.

Animals. Animals have always been important subjects for sculpture. Paleolithic man produced some extraordinarily sensitive animal sculptures both in relief and in the round. Representations of horses and lions are among the finest works of Assyrian sculpture. Egyptian sculptors produced sensitive naturalistic representations of cattle, donkeys, hippopotamuses, apes, and a wide variety of birds and fish. Ancient Chinese sculptors made superb small-scale animal sculptures in bronze and pottery. Animals were the main subject matter for the sculpture of the nomadic tribes of Eurasia and northern Europe, for whom they became the basis for elaborate zoomorphic fantasies. This animal art contributed to the rich tradition of animal sculpture in medieval art. Animals also served as a basis for semi-abstract fantasy in Mexican, Mayan, North American Indian, and Oceanic sculpture. The horse has always occupied an important place in Western sculpture, but other animals have also figured in the work of such sculptors as Giambologna, in the 16th century, and Antoine-Louis Barye, in the 19th, as well as numerous sculptors of garden and fountain pieces. Among modern sculptors who have made extensive use of animals or animal-like forms are Brancusi, Picasso, Gerhard Marcks, Germaine Richier, François Pompon, Pino Pascali, and François-Xavier-Maxime Lalanne.

Fantasy. In their attempts to imagine gods and mythical beings, sculptors have invented fantastic images based on the combination and metamorphosis of animal and human forms. A centaur, the Minotaur, and animal-headed gods of the ancient world are straightforward combinations. More imaginative fantasies were produced by Mexican and Mayan sculptors and by tribal sculptors in many parts of the world. Fantastic creatures abound in the sculpture produced in northern Europe during the early Middle Ages and the Romanesque period. Fantasy of a playful kind is often found in garden sculpture and fountains.

In the period following World War I, fantasy was a dominant element in representational sculpture. Among its many forms are images derived from dreams, the technological fantasy of science fiction, erotic fantasies, and a whole host of monsters and automata. The Surrealists have made a major contribution to this aspect of modern sculpture.

Other subjects. Architectural backgrounds in sculpture range from the simplified baldacchinos (ornamental struc-

GEKS

Narrative sculpture.
Lower portion of Trajan's Column, marble, Rome, c. AD 106–113. Height of one relief band about 127 cm.

tures resembling canopies used especially over altars) of early medieval reliefs to the 17th- and 18th-century virtuoso perspective townscapes of Grinling Gibbons. Architectural accessories such as plinths, entablatures, pilasters, columns, and moldings have played a prominent role both in Greek and Roman sarcophagi, in medieval altarpieces and screens, and in Renaissance wall tombs.

Outside the field of ornament, botanical forms have played only a minor role in sculpture. Trees and stylized lotuses are especially common in Indian sculpture because of their great symbolic significance. Trees are also present in many Renaissance reliefs and in some medieval reliefs.

Landscape, which was an important background feature in many Renaissance reliefs (notably those of Ghiberti) and, as sculptured rocks, appeared in a number of Baroque fountains, entered into sculpture in a new way when Henry Moore combined the forms of caves, rocks, hills, and cliffs with the human form in a series of large reclining figures.

There is nothing in sculpture comparable with the tradition of still-life painting. When objects are represented, it is almost always as part of a figure composition. A few modern sculptors, however, notably Giacomo Manzù and Oldenburg, have used still-life subjects.

NONREPRESENTATIONAL SCULPTURE

Kinds of nonrepresentational sculpture

There are two main kinds of nonrepresentational sculpture. One kind uses nature not as subject matter to be represented but as a source of formal ideas. For sculptors who work in this way, the forms that are observed in nature serve as a starting point for a kind of creative play, the end products of which may bear little or no resemblance either to their original source or to any other natural object. Many works by Brancusi, Raymond Duchamp-Villon, Jacques Lipchitz, Henri Laurens, Umberto Boccioni, and other pioneer modern sculptors have this character. The transformation of natural forms to a point where they are no longer recognizable is also common in many styles of primitive and ornamental art.

The other main kind of nonrepresentational sculpture, often known as nonobjective sculpture, is a more completely nonrepresentational form that does not even have a starting point in nature. It arises from a constructive manipulation of the sculptor's generalized, abstract ideas of spatial relations, volume, line, colour, texture, and so on. The approach of the nonobjective sculptor has been likened to that of the composer of music, who manipulates the elements of his art in a similar manner. The inclusion of purely invented, three-dimensional artifacts under the heading of sculpture is a 20th-century innovation.

Some nonobjective sculptors prefer forms that have the complex curvilinearity of surface typical of living organisms; others prefer more regular, simple geometric forms. The whole realm of three-dimensional form is open to nonobjective sculptors, but these sculptors often restrict themselves to a narrow range of preferred types of form. A kind of nonobjective sculpture prominent in the 1950s and '60s, for example, consisted of extremely stark, so-called primary forms. These were highly finished, usually coloured constructions that were often large in scale and made up entirely of plane or single-curved surfaces. Prominent among the first generation of nonobjective sculptors were Jean Arp, Antoine Pevsner, Naum Gabo, Barbara Hepworth, Max Bill, and David Smith. Subsequent artists who worked in this manner include Robert Morris, Donald Judd, and Phillip King.

DECORATIVE SCULPTURE

The devices and motifs of ornamental sculpture fall into three main categories: abstract, zoomorphic, and botanical. Abstract shapes, which can easily be made to fit into any framework, are a widespread form of decoration. Outstanding examples of abstract relief ornament are found on Islāmic, Mexican, and Mayan buildings and on small Celtic metal artifacts. The character of the work varies from the large-scale rectilinear two-plane reliefs of the buildings of Mitla in Mexico, to the small-scale curvilinear plastic decoration of a Celtic shield or body ornament.

Zoomorphic relief decoration, derived from a vast range of animal forms, is common on primitive artifacts and on Romanesque churches, especially the wooden stave churches of Scandinavia.

Decorative sculpture.
(Top) Detail of the patio of the Palace of the Columns, Mitla, Oaxaca state, Mexico, Mixtec culture, 9th–16th century. (Bottom) Norse wood carving, late 11th century, detail of doorway of Urnes stave church, Norway. Doorway 3.90 × 2.80 m.
(Top) Henri Lehmann

Botanical forms lend themselves readily to decorative purposes because their growth patterns are variable and their components—leaf, tendril, bud, flower, and fruit—are infinitely repeatable. The acanthus and anthemion motifs of Classical relief and the lotuses of Indian relief are splendid examples of stylized plant ornament. The naturalistic leaf ornament of Southwell Minster, Reims Cathedral, and other Gothic churches transcends the merely decorative and becomes superbly plastic sculpture in its own right.

SYMBOLISM

Sculptural images may be symbolic on a number of levels. Apart from conventional symbols, such as those of heraldry and other insignia, the simplest and most straightforward kind of sculptural symbol is that in which an abstract idea is represented by means of allegory and personification. A few common examples are figures that personify the cardinal virtues (prudence, justice, temperance, fortitude), the theological virtues (faith, hope, and charity), the arts, the church, victory, the seasons of the year, industry, and agriculture. These figures are often provided with symbolic objects that serve to identify them; for example, the hammer of industry, the sickle of agriculture, the hourglass of time, the scales of justice. Such personifications abound in medieval and Renaissance sculpture and were until recently the stock in trade of public sculpture the world over. Animals are also frequently used in the same way; for example, the owl (as the emblem of Athens and the symbol of wisdom), the British lion, and the American eagle.

Allegory and personification

Beyond this straightforward level of symbolism, the images of sculpture may serve as broader, more abstruse religious, mythical, and civic symbols expressing some of mankind's deepest spiritual insights, beliefs, and feelings. The great tympanums (the space above the lintel of a door

that is enclosed by the doorway arch) of Autun, Moissac, and other medieval churches symbolize some of the most profound Christian doctrines concerning the ends of human life and man's relations with the divine. The Hindu image of the dance of Śiva is symbolic in every detail, and the whole image expresses in one concentrated symbol some of the complex cosmological ideas of the Hindu religion. The Buddhist temple of Borobuḍur, in Java, is one of the most complex and integrated of all religious symbols. It is designed as a holy mountain whose structure symbolizes the structure of the spiritual universe. Each of the nine levels of the temple has a different kind of sculptural symbolism, progressing from symbols of hell and the world of desire at the lowest level to austere symbols of the higher spiritual mysteries at its uppermost levels.

In more individualistic societies, works of sculpture may be symbolic on a personal, private level. Michelangelo's "Slaves" have been interpreted as allegories of the human soul struggling to free itself from the bondage of the body, its "earthly prison," or, more directly, as symbols of the struggle of intelligible form against mere matter. But there is no doubt that, in ways difficult to formulate precisely, they are also disturbing symbols of Michelangelo's personal attitudes, emotions, and psychological conflicts. If it is an expression of his unconscious mind, the sculptor himself may be unaware of this aspect of the design of his work.

Many modern sculptors disclaim any attempt at symbolism in their work. When symbolic images do play a part in modern sculpture, they are either derived from obsolete classical, medieval, and other historical sources or they are private. Because there has been little socially recognized symbolism for the modern sculptor to use in his work, symbols consciously invented by individual artists or deriving from the image-producing function of the individual unconscious mind have been paramount. Many of these are entirely personal symbols expressing the artist's private attitudes, beliefs, obsessions, and emotions. They are often more symptomatic than symbolic. Henry Moore is outstanding among modern sculptors for having created a world of personal symbols that also have a universal quality; and Naum Gabo has sought images that would symbolize in a general way modern man's attitudes to the world picture provided by science and technology.

Symbolism of sculptural sites

Examples of sculpture of which the positioning, or siting, as well as the imagery is symbolic are the carved boundary stones of the ancient world; memorials sited on battlegrounds or at places where religious and political martyrs have been killed; the Statue of Liberty and similar civic symbols situated at harbours, town gates, bridges, and so on; and the scenes of the Last Judgment placed over the entrances to cathedrals, where they could serve as an admonition to the congregation.

The choice of symbolism suitable to the function of a sculpture is an important aspect of design. Fonts, pulpits, lecterns, triumphal arches, war memorials, tombstones, and the like all require a symbolism appropriate to their function. In a somewhat different way, the tomb sculptures of Egypt, intended to serve a magical function in the afterlife of the tomb's inhabitants, had to be images suitable for their purpose. These, however, are more in the nature of magical substitutes than symbols.

Uses of sculpture

The vast majority of sculptures are not entirely autonomous but are integrated or linked in some way with other works of art in other mediums. Relief, in particular, has served as a form of decoration for an immense range of domestic, personal, civic, and sacred artifacts, from the spear-throwers of Paleolithic man and the cosmetic palettes of earliest Egyptian civilization to the latest mass-produced plastic reproduction of a Jacobean linenfold panel (a carved or molded panel representing a fold, or scroll, of linen).

The main use of large-scale sculpture has been in conjunction with architecture. It has either formed part of the interior or exterior fabric of the building itself or has been placed against or near the building as an adjunct to it. The role of sculpture in relation to buildings as part of a townscape is also of considerable importance. Traditionally, it has been used to provide a focal point at the meeting of streets, and in marketplaces, town squares, and other open places—a tradition that many town planners today are continuing.

Sculpture has been widely used as part of the total decorative scheme for a garden or park. Garden sculpture is usually intended primarily for enjoyment, helping to create the right kind of environment for meditation, relaxation, and delight. Because the aim is to create a lighthearted arcadian or ideal paradisal atmosphere, disturbing or serious subjects are usually avoided. The sculpture may be set among trees and foliage where it can surprise and delight the viewer or sited in the open to provide a focal point for a vista.

Fountains, too, are intended primarily to give enjoyment to the senses. There is nothing to compare with the interplay of light, movement, sound, and sculptural imagery in great fountains, which combine the movement and sound of sheets, jets, and cataracts of water with richly imaginative sculpture, water plants and foliage, darting fish, reflections, and changing lights. They are the prototypes of all 20th-century "mixed-media" kinetic sculptures.

The durability of sculpture makes it an ideal medium for commemorative purposes, and much of the world's greatest sculpture has been created to perpetuate the memory of persons and events. Commemorative sculpture includes tombs, tombstones, statues, plaques, sarcophagi, memorial columns, and triumphal arches. Portraiture, too, often serves a memorial function.

Sculptural coins and medals

One of the most familiar and widespread uses of sculpture is for coins. Produced for more than 2,500 years, these miniature works of art contain a historically invaluable and often artistically excellent range of portrait heads and symbolic devices. Medals, too, in spite of their small scale, may be vehicles for plastic art of the highest quality. The 15th-century medals of the Italian artist Antonio Pisanello and the coins of ancient Greece are generally considered the supreme achievements in these miniature fields of sculpture.

Also on a small scale are the sculptural products of the glyptic arts—that is, the arts of carving gems and hard stones. Superb and varied work, often done in conjunction with precious metalwork, has been produced in many countries.

Finally, sculpture has been widely used for ceremonial and ritualistic objects such as bishop's croziers, censers, reliquaries, chalices, tabernacles, sacred book covers, ancient Chinese bronzes, burial accessories, the paraphernalia of tribal rituals, the special equipment worn by participants in the sacred ball game of ancient Mexico, processional images, masks and headdresses, and modern trophies and awards.

BIBLIOGRAPHY. EDWARD LANTERI, *Modelling and Sculpture: A Guide for Artists and Students*, 3 vol. (1965; previously pub. under the title *Modelling*, 3 vol., 1902–11), still an outstanding work on traditional methods; JACK C. RICH, *The Materials and Methods of Sculpture* (1947), comprehensive coverage of all except the most recent methods and materials; WILBERT VERHELST, *Sculpture: Tools, Materials, and Techniques* (1973), a wide-ranging survey with good coverage of modern materials; JOHN W. MILLS, *The Technique of Casting for Sculpture* (1967), and *Sculpture in Concrete* (1968), two useful technical handbooks; TREVOR FAULKNER, *The Thames and Hudson Manual of Direct Metal Sculpture* (1978), an informative work on a variety of historical and modern methods; UDO KULTERMANN, *The New Sculpture: Environments and Assemblages* (1968; originally published in German, 1967), a comprehensive account of these two recently developed forms of sculpture; RUDOLF WITTKOWER, *Sculpture: Processes and Principles* (1977), an authoritative account of the interaction of techniques and aesthetics in the history of sculpture; L.R. ROGERS, *Sculpture* (1969), and *Relief Sculpture* (1974), two books dealing with the principles and techniques of sculpture and their bearing on its appreciation as an art form.

(L.R.R.)

The History of Western Sculpture

Sculpture may be broadly defined as the art of representing observed or imagined objects in solid materials and in three dimensions. Like Western painting, Western sculpture has tended to be humanistic and naturalistic, concentrating upon the human figure and human action studied from nature. Early in the history of the art there developed two general types: statuary, in which figures are shown in the round, and relief, in which figures project from a ground.

Western sculpture in the ancient world of Greece and Rome and from the late Middle Ages to the end of the 19th century twice underwent a progressive development, from archaic stylization to realism; the term progressive here means that the stylistic sequence was determined by what was previously known about the representation of the human figure, each step depending upon a prior one, and not that there was an aesthetic progression or improvement. Modern criticism has sometimes claimed that much was lost in the change. In any event, the sculptors of the West closely observed the human body in action, at first attempting to find its ideal aspect and proportions and later aiming for dramatic effects, the heroic and the tragic; still later they favoured less significant sentiments, or at least more familiar and mundane subjects.

The pre-Hellenic, early Christian, Byzantine, and early medieval periods contradicted the humanist-naturalist bias of Greece and Rome and the Renaissance; in the 20th century that contradiction has been even more emphatic. The 20th century has seen the move away from humanistic naturalism to experimentation with new materials and techniques and new and complex imagery. With the advent of abstract art, the concept of the figure has come to encompass a wide range of nonliteral representation; the notion of statuary has been superseded by the more inclusive category of freestanding sculpture; and, further, two new types have appeared: kinetic sculpture, in which actual movement of parts or of the whole sculpture is considered an element of design; and environmental sculpture, in which the artist either alters a given environment as if it were a kind of medium or provides in the sculpture itself an environment for the viewer to enter.

This article is divided into the following sections:

European Metal Age cultures 60
- Aegean and Eastern Mediterranean 60
  - The Early Bronze Age (3000–2000 BC)
  - The Middle Bronze Age (2000–1600 BC)
  - The Late Bronze Age (1600–1100 BC)
- Western Mediterranean 63
  - Bronze Age cultures
  - Iron Age cultures

Ancient Greek 65
- The Geometric period 66
- The Orientalizing period 66
- The Archaic period 66
- The Classical period 67
  - Early Classical (*c.* 500–450 BC)
  - High Classical period (*c.* 450–400 BC)
  - Late Classical period (*c.* 400–323 BC)
  - Hellenistic period

Roman and Early Christian 72
- The last century of the Republic 73
- The empire 75
- Early Christian 81

The Middle Ages 82
- Eastern Christian 82
  - Constantinople and the Byzantine Empire
  - Georgia
  - Armenia
  - Coptic Egypt
- Western Christian 86
  - Carolingian and Ottonian periods
  - Romanesque

Gothic 88
- Early Gothic 88
- High Gothic 90
  - Italian Gothic
  - International Gothic
- Late Gothic 93

The Renaissance 95
- Italy 95
  - Early Renaissance
  - High Renaissance and Mannerism
- Mannerist sculpture outside Italy 99

The Baroque period 99
- Italy 99
  - Early and High Baroque
  - Late Baroque
- Baroque and Rococo outside Italy 102
  - Spain
  - Flanders
  - France
  - England
  - Central Europe
  - Russia
  - Latin America

Neoclassical and Romantic sculpture 104
- Neoclassicism 104
  - "Decorum" and idealization
  - Relation to the Baroque and the Rococo
- 19th-century sculpture 107

Modern sculpture 108
- 19th-century beginnings 108
- The 20th century 108
  - Avant-garde sculpture (1909–20)
  - Constructivism and Dada
  - Conservative reaction (1920s)
  - Sculpture of fantasy (1920–45)
  - Other sculpture (1920–45)
  - Developments after World War II

Bibliography 113

European Metal Age cultures

AEGEAN AND EASTERN MEDITERRANEAN

Cultures of the Aegean

Aegean civilization is a general term for the prehistoric Bronze Age cultures of the area around the Aegean Sea covering the period from *c.* 3000 BC to *c.* 1100 BC, when iron began to come into general use throughout the area. From the earliest times these cultures fall into three main groups: (1) the Minoan culture (after the legendary king Minos) of Crete, (2) the Cycladic culture of the Cyclades islands, and (3) the Helladic culture of mainland Greece (Hellas). For convenience, the three cultures are each divided into three phases, Early, Middle, and Late, in accordance with the phases of the Bronze Age. The culture of Cyprus in the eastern Mediterranean, although it commenced somewhat later than those of the Aegean, came to parallel them by the Middle Bronze Age. The Late Bronze Age phase of the mainland is usually called Mycenaean after Mycenae, the chief Late Bronze Age site in mainland Greece.

The first centre of high civilization in the Aegean area, with great cities and palaces, a highly developed art, extended trade, writing, and use of seal stones, was Crete. Here from the end of the 3rd millennium BC onward a very distinctive civilization, owing much to the older civilizations of Egypt and the Middle East but original in its character, came into being.

The Cretan (Minoan) civilization had begun to spread by the end of the Early Bronze Age across the Aegean to the islands and to the mainland of Greece. During the Late

Bronze Age, from the middle of the 16th century onward, a civilization more or less uniform superficially but showing local divergences is found throughout the Aegean area. Eventually people bearing this civilization spread colonies eastward to Cyprus and elsewhere on the southern and western coasts of Asia Minor as far as Syria, also westward to Tarentum in southern Italy and even perhaps to Sicily. In the latter part of this period, after about 1400 BC, the centre of political and economic power, if not of artistic achievement, appears to have shifted from Knossos in Crete to Mycenae on the Greek mainland. (Ed.)

The Early Bronze Age (3000–2000 BC). *Early Minoan.* The early Minoan period saw a thousand years of peaceful development, which eventually gave place to the full flowering of the Minoan spirit, the Middle Minoan period. Pottery was preeminent among the Early Minoan arts.

Early Cycladic. The Early Cycladic culture developed on parallel lines to the Early Minoan. Thanks to obsidian from Melos, marble from many islands, and local sources of gold, silver, and copper, the Cycladic islanders rapidly became prosperous. As in Crete, the Early Bronze Age merged without incident into the Middle Bronze Age.

The Early Cycladic period is celebrated principally for its statuettes and vases carved from the brilliant coarse-crystalled marble of these islands. The statuettes, mostly of goddesses, are among the finest products of the Greek Bronze Age. They owe their charm to the extreme simplification of bodily forms. The typical "Cycladic idol" is a naked female, lying with her head back, her arms crossed over her breasts. These figures vary in size from a few inches to more than six feet in length (Figure 1).

Early Helladic and Early Cypriot. Mainland Greece probably received its Bronze Age settlers from the Cyclades, but the two cultures soon diverged. A prosperous era arose about 2500 BC and lasted until about 2200. Sculpture was overshadowed by pottery, metalwork, and architecture among the Early Helladic arts. In the Early Cypriot, the only surviving sculptures are a series of steatite cruciform figures of a mother goddess (3000–2500 BC) stylized in much the same way as contemporary Cycladic idols, from which they may have been derived.

The Middle Bronze Age (2000–1600 BC). *Middle Minoan.* The Middle Minoan period differs principally from the Early Minoan in the creation of palaces and a palatial life and art. Large-scale sculpture seems not to have found much favour in Crete, although fragments of life-size figures from this period were discovered in the Cyclades in the late 20th century. Miniature sculpture of the highest quality, some of it of fired sand and clay, was produced

Figure 1: Marble Cycladic idol from Amorgos, Greece, *c.* 2500 BC. In the National Archaeological Museum, Athens.
Emile Serafis

from at least as early as 1700 BC. Good examples are two female figures (called "Snake Goddesses") from Knossos, dated about 1700 BC (Archaeological Museum, Iráklion, Crete). These women stand with their arms in front of them, holding sacred snakes; they wear a flounced skirt and tight belt, and their breasts are bare.

Minoan "Snake Goddess" figures

Middle Cycladic, Middle Helladic, and Middle Cypriot. During the Middle Cycladic period, the Cyclades suffered a diminution in prosperity and seem to have become polit-

Alison Frantz

Figure 2: *Carved stone Minoan vessels of the Late Bronze Age (1600–1100 BC).* (Left) Serpentine rhyton (drinking vessel) in the form of a bull's head, steatite with gold-plated horns (now restored), from the Little Palace at Knossos, Crete, *c.* 1500 BC. In the Archaeological Museum, Iráklion, Crete. Height without horns 20.6 cm. (Right) Extant portion of the "Harvester Vase," from Ayia Triádha, steatite, *c.* 1600 BC. In the Archaeological Museum, Iráklion. Diameter 14 cm.

ically subordinate to Crete. Two waves of Indo-European peoples seem to have descended on the Greek mainland, one about 2200 BC and the other about 2000 BC. They destroyed much and for long contributed little to Greece's artistic heritage. The pottery of this period, however, is of high quality. The Middle Cypriot period was a development of the Early Cypriot. As on the mainland, no important art apart from pottery has survived.

The Late Bronze Age (1600–1100 BC). *Late Minoan.* Prosperity and artistic achievement remained at a high level until about 1450 BC, when all the great centres of Cretan culture were destroyed by earthquakes (probably connected with a cataclysmic eruption of the volcanic island of Thera). After these disasters, only the palace at Knossos was restored for occupation. About 1375 BC, however, the palace at Knossos was destroyed by fire. Thereafter Crete was a second-class power and became somewhat of a cultural backwater. Miniature sculpture was still popular. No longer in faience, figures were increasingly made of bronze, ivory, and terra-cotta. Some of the bronzes, cast solid by the "lost wax" process (using a wax model), are very fine, the earliest being the best. The subjects include male worshippers wearing boots, tight belt, and kilt; women (perhaps goddesses) dressed like the faience snake goddesses of the Middle Minoan period; and animals, especially bulls.

Carved-stone vases

Carved-stone vases were made between 1600 and 1450 BC. Elegant vessels were carved from such diverse materials as marble, obsidian, and steatite (Figure 2). Others, of soft stone, were made in the shape of bulls' heads, astonishingly true to life, or were carved in relief, with religious or court ritual scenes, and covered with gold leaf.

The art of the seal engraver flourished until 1375 BC. Religious subjects, scenes of the bullring, and depictions of animals in their natural setting were popular. Even the exaggerations of the style reflect careful observation of the movements of the animals and their idiosyncratic anatomy, but they also relate the forms depicted to the shape of the stone—the curve of a bull's back or horns to that of the edge, for instance (Figure 3).

Mycenaean. Mainland Greece enjoyed renewed contacts with Crete *c.* 1600 BC, and a rich culture, based on the Late Minoan, rapidly came into being. The Mycenaeans gained control of Crete *c.* 1450 BC, and between 1375 and 1200 BC they became masters of an empire that stretched from Sicily and southern Italy in the west to Asia Minor and the Levant coast in the east. About 1200 BC, however, many of the Mycenaean strongholds were destroyed by fire. There were signs of a renaissance, but the end of Mycenaean civilization came *c.* 1100 BC.

Figure 3: Impression of a seal stone from Vapheio, Greece, dating from *c.* 1500 BC.

From *Crete and Mycenae* published by Thames & Hudson, London, and Harry N. Abrams, New York; photograph, Hirmer Fotoarchiv, Munchen

The Mycenaeans seem to have had more of a taste for monumental sculpture than had their Minoan mentors. Of the few surviving examples, the best known is a relief over the Lion Gate at Mycenae (*c.* 1250 BC), in which two lions confront each other across an architectural column (Figure 4). Probably heraldic in concept, this design is comparable with those on tiny seals and ivories of Cretan inspiration. Sculpture on a small scale, in ivory, bronze, and terra-cotta, generally Minoan in character, remained popular.

The Lion Gate at Mycenae

Late Cypriot. Cyprus reached its highest degree of prosperity in the Late Cypriot period, due to increased exploitation of its copper mines. There were close commercial relations not only with the Levant coast, as before, but also with Egypt, Crete, and Mycenaean Greece (the latter being close from 1400 BC). About 1200 BC Mycenaean Greeks, refugees from their homeland, settled in Cyprus. They introduced their skills and produced many luxury articles in a mixed Mycenaean-Cypriot style. Cyprus escaped the invasions that finally destroyed Mycenaean and Minoan culture, but its own culture did not last much longer. By 1050 BC, for reasons that are not clear, it, too, had ceased to exist.

From *Crete and Mycenae* published by Thames & Hudson, London, and Harry N. Abrams, New York; photograph, Hirmer Fotoarchiv, Munchen

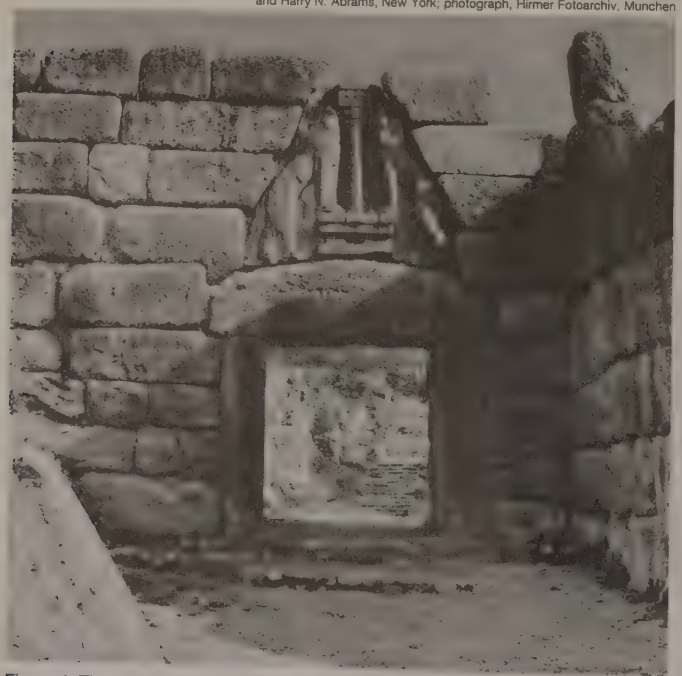


Figure 4: The Lion Gate at Mycenae, Greece, *c.* 1250 BC.

As in Crete, large-scale sculpture was rejected in favour of small-scale work. A bronze figure of a horned god (shortly after 1200 BC) from Enkomi (Cyprus Museum, Nicosia) shows a successful blend of Mycenaean and Cypriot elements. A good example of these characteristics is a carved ivory gaming box (British Museum), also from Enkomi, whose style shows a blend of Mycenaean and Middle Eastern motifs. (R.A.Hi./Ed.)

WESTERN MEDITERRANEAN

Like central and northern Europe, although to a lesser degree, the western Mediterranean was considerably behind the eastern Mediterranean, where civilization, the arts, and writing were born much earlier. The development of the metallurgical industry did not occur simultaneously in the various regions of the western Mediterranean, but it did bring important innovations in the mode of living and, of course, in the arts.

The Chalcolithic (Copper-Stone) era began in Spain at the end of the 3rd millennium BC at Los Millares, near Almería, and in Italy at the beginning of the 2nd millennium with the Remedello civilization. Bronze appeared not long afterward, around 1800 BC, in Italy and Sardinia. The Bronze Age in Italy gave way to the Iron Age at the beginning of the 1st millennium BC, but elsewhere, as in Sardinia or Spain, it lasted longer. The Iron Age flourished on the Illyrian coasts and in Italy from 900 to 800 BC; it also lasted varying lengths of time according to locale. After this, one may speak of the civilizations of Magna Graecia, of Rome, or of Etruria.

During the metal ages, popular migrations, commerce, and wars increased, which resulted in the rise of cities and of fortified works for their protection and defense, such as the talayots (round or quadrangular towers) of the Balearic Isles and the nuraghi (round towers) of Sardinia. With respect to the plastic arts, one particularly remarkable phenomenon was the birth and multiplication of megalithic human representations, which gained in number and importance from the 3rd to the 1st millennium BC. The Neolithic monuments, menhirs (single, vertical megaliths) and dolmens (structures of two vertical stones capped by a horizontal one), which had arisen in the megalithic era, continued to appear in the Copper and Bronze Ages, but then—here and there in Spain, Sardinia, Corsica, Liguria, and in the south of France—stelae-menhirs (carved or inscribed stone slabs used for commemorative purposes), like the stammerings of Western figure sculpture, imitated the human form. They maintained certain stylistic relations with rock engravings of mountainous regions, such as the Val Camonica.

Human representations

Bronze Age cultures. *Sardinia and Corsica.* The nuraghic civilization had an original sculpture expressed in a large production of bronze statuettes, about 500 of which have been found in nuraghi, temples, houses, and tombs (Figure 5). These figurines represent all classes of the proto-Sardinian populations—military chiefs, soldiers, priests, and women, as well as heroes and gods—in what seems to the modern viewer to be an engagingly direct but also sophisticated geometric style. The greatest number of these bronzes are today in the Museo Archeologico Nazionale in Cagliari, Sardinia. Some have been discovered in Etruscan tombs of Vetulonia and Vulci and have been dated to the period extending from the 9th to the 6th century BC.

Corsican menhir, or stela, statuary

Corsican menhir, or stela, statuary constitutes a group of special interest. The stone is imbued with life by a sculptural art that involves roughing-in of the head, animation of the upper portion of the body, and placement of a few elements of ornamentation or weaponry (sculpted in relief or, more rarely, engraved) on the schematically anthropomorphic image. These primitive statues are masculine and, no doubt, represent family or tribal heads made heroic or divine. This megalithic stela statuary art appears not only on Corsica but also in various other countries and regions of the western Mediterranean, including Spain, Sardinia, Liguria, and, in southern France, Provence, Aveyron, Hérault, and Gard, though to a lesser degree. The advance of this type of megalithic sculptural art is difficult to follow, but it is clear that these different groups are related, with close affinities existing between the stelae-menhirs of Corsica and those of the Ligurian coast. Such art is everywhere the expression of a patriarchal society seeking to impose on men's vision, massively and not without grandeur, the image of the departed ancestors.

Figure 5: Bronze statuette, nuraghic civilization, from an unknown site in Sardinia, c. 7th century BC. In the Nationalmuseet, Copenhagen. Height 20.3 cm.

By courtesy of the Nationalmuseet, Copenhagen

Italy. From the Bronze Age of far northern Italy there survives an exceptional collection of rock engravings, a remarkable extension of an art that, in fact, had been represented in the prehistoric era and had not yet vanished completely. About 20,000 rock engravings have been found between altitudes of 5,000 and 5,600 feet (1,500 and 1,700 metres) in the Val Camonica, north of the town of Brescia. This art is found again further west, in the Maritime Alps of France on Monte Bego, between altitudes of 6,600 and 8,900 feet, and less remarkably elsewhere. What is exceptional about the carvings of the Val Camonica is that they represent a variety of subjects—rituals, battles, hunting, and daily labour—and that these were treated as compositions.

Rock engravings

Although engraving played a minor role in the case of the menhir statuary mentioned earlier, relations do exist between the sculpted works and the Camunian images of Monte Bego. The same representations of collar torques appear on the menhir statuary of Gard, Aveyron, and Tarn, on the one hand, and on certain monumental engravings of the Val Camonica, on the other. Some kind of relationship thus unites the arts of rock engraving and stela statuary in the Bronze Age.

Iron Age cultures. *Italy.* The Italian peninsula, which in the Bronze Age had been only one among many centres of civilization, took on a special importance in the Iron Age. Widespread and powerful cultural and artistic centres grew up there, first in the Villanovan civilization and later in the Etruscan; their influence was disseminated into the surrounding areas.

At the beginning of the 1st millennium BC there began to develop in the Po plain, in Tuscany, Latium, and some areas of Lucania, a new cremating civilization, which draws its name from that of the Villanova necropolis, discovered near Bologna. It is obviously related to the so-called Urnfield civilization that, at the end of the Bronze Age and beginning of the Iron Age, extended over central and eastern Europe and had developed a metal art with geometric and abstract ornamentation. The ashes of the dead were placed in urns thrust in level with the soil. From the Urnfield civilization arose two others: the Hallstatt civilization, which spread into the Balkans, northern and

Villanovan civilization

central Europe, and France, beyond the Pyrenees; and in Italy the Villanovan civilization and the civilizations that, to the east and west of the Po plain, were related to it, the so-called Golasecca civilization in the great lakes region and the Este civilization in the Venice area.

These Italic civilizations of the Early Iron Age, which appeared at the beginning of the 1st millennium BC and lasted for varying lengths of time, multiplied the number of dwelling sites. Originating as outposts established on naturally strong positions, they began to resemble towns as population increased.

The cinerary urn, which was made first of terra-cotta and later of bronze, assumes, by its form, a symbolic value (Figure 6). Biconical in form and covered with an

Figure 6: Villanovan canopic vase from Chiusi, Italy, terra-cotta, 6th century BC. In the Museo Archeologico, Florence. Height 44 cm.
By courtesy of the Archeological Museum, Florence

overturned cup, later with a helmet, it schematically represents the appearance of the human body. Sometimes, as in examples from Latium and Tuscany, the funerary vessel is in the form of a hut or cabin—the house of the dead person whose remains it holds. The ornamentation, painted or engraved on the vases and engraved or in relief on metal objects, is in a geometric, nonfigurative style. Human or animal forms appear only rarely—in the decoration of small utilitarian objects such as vase handles and horse bits. It is a severe art, therefore, which essentially limits itself to linear exercises. Even motifs such as the disk, the solar boat, and the birds that encircle them, inherited from a more distant past and possessing primitively religious value, take on a stylized air and become abstract figures.

A naturalistic note is provided by the imagery that decorates, in zones of superimposed relief, bronze vessels called situlae, a kind of pail found in Eastern countries and in the eastern Alps. These situlae were made in Venetian workshops in particular and were very popular in the neighbouring areas. They rapidly underwent an Etruscan influence, however, that tended to give prominence in the chased ornamentation to human figures at feasts, games, or funerals, as in the masterpiece known from the place of its discovery as the Certosa Situla (Museo Civico Archeologico, Bologna). (R.Bl./Ed.)

Etruria, Latium, and the Faliscan districts fall into three main areas of artistic production: northern, central, and southern, each centred upon cities with a distinctive artistic style. In the southern areas the chief centres were Caere and Veii, in which the Etruscan style most closely approached the Greek. In central Etruria, Vulci was evidently the leading art centre, although Tarquinia was unsurpassed in the beauty of its wall paintings. There were several potteries in Vulci, and the greater part of the central Etruscan bronzes, artistically the best, were produced there. The north was dominated by Clusium, although Perugia seems to have been important along with lesser centres at Volterra and Fiesole.

The very earliest examples of Etruscan statuary are flat, rectilinear figurines from Vetulonia and Capodimonte di Bolsena. These figures occur in later contexts in the Regolini-Galassi and Bernardini tombs, both of which contain pieces in a more advanced style that cannot have developed much later. These are statuettes of women with pigtails and long skirts depicted in a manner that suggests a north Syria influence, although this female type, frequently copied in ivory and amber, is certainly of local origin.

The earliest evidence of Greek influence is the presence of centaurs, perhaps transmitted on Corinthian vases. Their style in Etruria is Orientalizing, with a slim body and elongated legs, perhaps reflecting Cretan influence. These and other mythical creatures found great favour with the Vulci stonemasons. To archaic works of early Etruscan sculpture certain Greek parallels can be found in the late 7th and early 6th centuries, and in general characteristics the works still followed the Greek Archaic Daedalic tradition. The next change in style took place *c.* 550 BC, when art became distinctively Ionian. These new influences can be seen earliest in such pieces of bronze work as the Loeb Tripod from San Valentino near Perugia and the Monteleone chariot platings (in the Metropolitan Museum, New York City), but they soon become apparent also in the relief designs on *bucchero pesante* (heavily embossed black pottery) and in architectural reliefs like those from Tarquinia. By the end of the 6th century BC Veii possessed an excellent school of terra-cotta sculptures in Ionian styles. The statues of Apollo and of a votaress suckling a child are elaborately stylized in features, draperies, and muscles. Clay statuary, still retaining traces of former painting, was made in many Etruscan centres. Examples in the more mature classical style that began in the last quarter of the 5th century are the satyr-and-maiden groups from Satricum (modern Conca) in the Museo Nazionale di Villa Giulia, Rome, which contains a rich collection of architectural terra-cottas from Caere, Falerii, Veii, Satricum, and other sites.

Ionian influences

These pieces of statuary were designed to stand on temple roofs, and the socketed bases by which they were fixed have survived. Terra-cotta sculpture was also used for antefixes for these temples but above all for funerary sculpture. Sarcophagi with the sculptured figures of the husband and his wife reclining on the lids seem to have begun late in the 6th century, the date of the haunting sarcophagus from Caere (Villa Giulia, Rome). Bronze sculptures were also produced from the end of the 6th century, beginning with the famous She-Wolf, the symbol of modern Rome (Musei Capitolini, Rome), and the later Chimera from Arezzo (Museo Archeologico, Florence) or the so-called Mars of Todi (Vatican Museums) of the early 4th century BC.

In spite of great achievements in sculpture in the round, most of what has survived is in low relief, and a series of fine 6th–5th-century relief sarcophagi from Clusium, depicting dances, funeral games and banquets, or the journey of the dead to the underworld, are a major source of information on Etruscan everyday life. Superbly carved gravestones of the late mid-6th century are known from Clusium and Settimello, but the disk- and horseshoe-shaped gravestones of the Bologna, Fiesole, and Populonia graves have crude reliefs. (W.Cu./Ed.)

Sculpture developed but did not seek, as in Greece, to represent the idealized body of athletes and gods, attempting instead to represent the figure and features of the deceased (Figure 7). There was a continuing taste for real or fantastic animals such as lions, panthers, and sphinxes, and the Etruscan imagination seems to have been haunted by these beasts and demons, the vigilant guardians of the tombs.

Iberia. Whether in the form of great statuary or small votive images, Iberian figurative art was essentially religious and intended to represent sacred animals, deities, and their worshippers. Although much influenced by Greek and other sources, these works are vigorous and original, as may be seen from "La dama de Elche" (Figure 8) and "La dama de Cerro de los Santos" in the Museo Arqueológico Nacional at Madrid. In the latter, a hieratic

Figure 7: Detail of "Reclining Couple," clay sarcophagus from Caere, Italy, late 6th century BC. In the Museo Nazionale di Villa Giulia, Rome. Length of entire sarcophagus 2.01 m.
Alinari—Art Resource

visage, with a severity not unlike some of the ideal heads of classical Greece, is adorned with a superabundance of heavy Iberian jewels. (R.Bl./Ed.)

Ancient Greek

Greek art no doubt owed much indirectly to the Minoan-Mycenaean civilization (now known in its later stages to have been Greek), which disintegrated at the end of the 2nd millennium BC, partly under the impact of a series of invasions from the Balkans. The period covered by this section, however, begins about 900 BC with the kaleidoscopic rearrangement of invaders and earlier inhabitants into a new pattern, which was followed by a steady artistic development—continuing without interruption down to the conquest of Greece by Rome in 146 BC. Even this diverted, rather than interrupted, the flow, and Greek artists continued to be predominant under the Roman Empire and beyond that into the Byzantine. But after Greece had become a Roman province, Greek art fell increasingly under the patronage of Romans and was devoted either to expressing Roman ideals or to reproducing older works of art. It is therefore reasonable to regard the later years of the 1st century BC, when the Roman Empire was forming, as the later limit of the period.

Five stages of Greek art

Within this period it is convenient to distinguish five stages of development. Their names are modern and arbitrary; the divisions between them are not equally sharp and do not apply equally to all parts of the Greek world, but they serve as a general guide to successive trends.

The first is the Geometric period (so-called from the rectilinear character of its art) from about 900 to about 800 BC, when Greece was self-contained and contact with the outside world was rare.

The second, the Orientalizing period, for about a century and a half from 800 BC, is one of contact with the East, a contact that had been broken by the upheavals at the end of the 2nd millennium.

The third period, the Archaic, from about 650 to about 480 BC, is characterized by the gradual absorption of Oriental elements and the rise and development of archaic Greek art.

The fourth period, from about 480 to about 330 BC, is known as the Classical; its beginning is marked by the rise of the sculptors Myron, Phidias, and Polyclitus and the painter Polygnotus, and its end, by the work of Scopas, Praxiteles, and Lysippus. (The word classical, which originally meant simply first-class, can also be used either in a narrower sense than this to denote only the Phidian age—*i.e.*, 50 years in the middle of the 5th century BC—or in a broader sense to cover the whole of post-Mycenaean Greek art from Geometric to late Roman.)

The fifth period is the Hellenistic, from about 330 BC, when the conquests of Alexander the Great opened new areas to the Greeks and the division of his kingdom among his Greek successors after his death in 323 diffused Greek art over the greater part of the known world, down to the late 1st century BC. Hellenistic symbolism and Hellenistic technical skill continued as living traditions under the Romans.

Statues were of limestone, marble, bronze, gold and ivory, terra-cotta, and wood. After the Archaic period the use of wood and of limestone seems to have been rare, as was the use of terra-cotta for statues of large size, although it should be noted that sculpture in the first and last of these materials tended to be ephemeral. The group of Orpheus and the two harpies that was restored at the J. Paul Getty Museum, Malibu, California, in the 1980s is astonishing not only for its quality but also for its size, and yet many other such figures may have been produced. Full-size statues of gold and ivory were rare at all times because of their cost; statues with gilded wooden bodies and marble extremities were sometimes made instead. For statuettes, ivory and amber, limestone, marble, wood gold, silver, bronze, and terra-cotta were used; of these, terra-cotta was by far the most common, bronze and marble less so, and the rest rare. Extremely valuable because they can often be dated with accuracy are the types of sculpture used for the decoration of buildings: acroteria (*i.e.*, figures on the tops or ends of gables); figures in the low triangular field of the pediment under the gable (both of these are usually almost in the round); sculptured panels (metopes) of the Doric frieze, which are usually in high or very high relief; and the continuous Ionic frieze, which is usually in low relief.

Hollе Bildarchiv, Baden-Baden, West Germany

Figure 8: "La dama de Elche," painted limestone bust from Elche, Alicante, Spain, 5th century BC. In the Museo Arqueológico Nacional, Madrid. Height 56.0 cm.

Sources of modern knowledge of Greek sculptures

Of the many thousands of statues produced during the period in which Greek art flourished, not more than a few dozen survive, and those mostly mutilated. Knowledge of the history of Greek sculpture depends partly on these and partly on the architectural sculptures—both of high importance, since they are original. Much can also be learned about the general development of sculptural style from the small bronzes, often of very high quality, and from the terra-cottas. Of the small bronzes many, and of the terra-cottas very many, have survived, but they were made by independent artists and did not copy contemporary statues closely. The great bulk of evidence comes from copies made by Greeks, for Roman patrons, of originals now destroyed. Such evidence is invaluable but not entirely reliable. There is also literary evidence, but much of this is also second-hand or dates from long after the period in which the sculptures in question were made.

THE GEOMETRIC PERIOD

In the 9th century BC Greece was settling down again after upheavals and migrations both into and out of the mainland. It seems that invaders from the north brought with them the germs of an artistic style that developed into the Greek Geometric tradition.

In addition to the pottery, the Geometric period produced some terra-cottas and many small bronzes. The bronzes tended to be flat at first but became more solid and less angular as casting direct from wax models superseded cutting from bronze plates. Birds and other animals, especially horses, were popular and often admirably done; men, perhaps because their form commanded less imaginative interest, were not so successfully rendered; in the later stages of geometric art, groups of some complexity were attempted—a doe with her fawn, a man fighting (or greeting) a centaur, even a lion hunt complete with dogs. (B.As./Ed.)

THE ORIENTALIZING PERIOD

Sculpture of the Orientalizing period was profoundly affected by technical and stylistic influences from the East. In about 700 BC, the Greeks learned from their Eastern neighbours how to use molds to mass-produce clay relief plaques. Widely adopted, this technique helped to establish in Greece a stereotyped convention for figure representation, even in freestanding, unmolded sculptures; and a strong Eastern stylistic influence ensured that the convention was Oriental in flavour—in most cases a frontal pose with stiff patterned hair and drapery rendered in a strictly decorative manner. The adoption of this convention, which has come to be known as Daedalic style (after Daedalus, the legendary craftsman of Crete, where the style especially flourished), put an end to the development of naturalism and freedom in miniature sculptures that had shown promise in the Geometric period, and eventually became representative of even major Greek sculpture in the mid-7th century BC.

The Daedalic style

Figure 10: "The Calf Bearer," marble statue, *c.* 570 BC. In the Acropolis Museum, Athens. Height 1.65 m.

Hirmer Fotoarchiv, Munchen

By courtesy of the Metropolitan Museum of Art, New York, Fletcher Fund, 1932

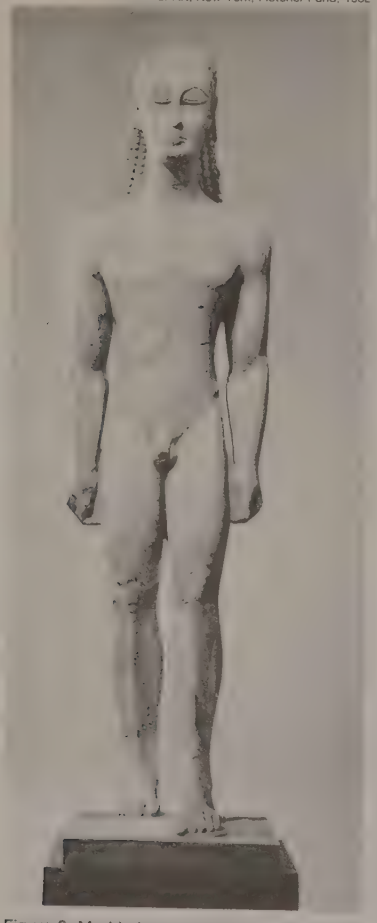


Figure 9: Marble kouros, *c.* 600 BC. In the Metropolitan Museum of Art, New York City. Height 1.86 m.

In about 640 BC, however, a second Eastern influence began to be felt. As with the gigantic architecture of Egypt, the Greeks were impressed with the monumentality of Egyptian statuary, larger than life-size and executed in hard stone instead of the limestone, clay, or wood to which the Greeks had been accustomed. The Greeks learned the techniques of handling the harder stone in Egypt, and at home they turned to the fine white marble of the Cyclades islands (mainly Paros and Naxos) for their materials. It was at this time that the first truly monumental examples of Greek sculpture appeared. The idiom and proportions were at first still Daedalic. By about 630 BC, however, first in the islands and later in mainland Greece, they were carving freestanding figures of naked men that were copies of types formerly seen only in minor art and that owed something in proportion and details of pose to the common Egyptian standing figures. This new series of life-size or larger marble youths (kouroi) reveals rapid developments in technique and style, notably a transition from the Daedalic past to greater naturalism through the new monumental manner (Figure 9). The earliest of these figures were, as might be expected, dedications in sanctuaries, especially on the island of Delos, but some were grave markers, as on another island, Thera. At the same time, the older style was used for relief decoration of temples in Crete and Greece, particularly at Mycenae.

The kouros figure

THE ARCHAIC PERIOD

The kouroi, which had become standardized as freestanding statues of naked youths with hands to sides and one leg advanced, were the most representative examples of Archaic sculpture (Figure 9). At first their proportions were based on theory rather than observation; much the same was true of the anatomical details, which were treated as separate patterns applied to the figure with no proper understanding of their physiological relationships. Growing awareness of natural forms, although still without systematic study of the model, together with technical

The Archaic kouros and kore

mastery, led to a realism that is striking in comparison with the Daedalic pieces of the Orientalizing period (Figure 10). Still, the overriding considerations of proportion and pattern were never subordinated to nature. Only in the years just before the Persian invasion of 480 BC did some sculptors recognize the organic structure of the body and succeed in showing a truly relaxed pose, with the weight shifted onto one leg and the hips and torso consequently tilted to break the rigid symmetry of the characteristic kouros of the Archaic period (Figure 11).

In the female counterpart of the kouros, the kore, Archaic sculptors were again preoccupied with proportion and pattern—the pattern of drapery rather than of anatomy (Figure 12). Ionian (Chios, Samos) and island (Naxos) sculptors took the lead in developing decorative schemes for rendering the fall and splay of the folds of the loosely draped Ionic dress (chiton) and overmantle (himation). These patterns, like the anatomy of the kouroi, suggest nature rather than copy it; the strict logic of dressmaking is never observed by the sculptor, who uses the natural gesture of pulling a long skirt up and to one side first to produce a pleasing pattern of folds and only later to reveal the contours of the legs and body beneath. Most of the korai, like the kouroi, stood as dedications in sanctuaries, the richest series being from the Acropolis at Athens (these were overthrown by the Persians and then piously buried by the returning Athenians). Few of these statues were grave markers.

Archaic architectural sculpture

In the addition of sculpture to architecture, the determining factor was usually its position on the building. On a Doric temple, for instance, the metope frieze offered a series of rectangular plaques for reliefs that could accommodate two or three figures. There was a tendency in the Archaic period to let the action run on from one metope to the next, regardless of the intervening triglyph, a practice that was later abandoned. Above the frieze, the pediments formed by the gabled roof provided an awkward field—a long, low triangle. The sculptors of early temple pediments met the problem by depicting separate groups of different sizes (Figure 13), as at Corcyra (Archaeological Museum, Kérkira, Greece), or by devising monster bodies to fill the shallow corners, as in Athens (Acropolis Museum, Athens). Later, the advantages of using fighting groups with falling and fallen bodies were discovered; this type is represented at Athens and Aegina (Munich). The later Archaic pedimental figures were executed virtually in the round,

Figure 12: Marble statue of a woman dedicated by Cheramyes to Hera, found in the Heraeum on Samos, Greece, c. 560 BC. In the Louvre, Paris. Height 1.92 m.

Archives Photographiques

standing against or just free from the background of the gable. Because these figures, unlike the kouroi and korai, were often in violent action, it may have been through meeting the problems of architectural sculpture that the artist arrived at a better understanding of the dynamics of the human body.

Work in relief also was used on gravestones, chiefly in Athens, for decorative bases of columns and for the frieze decoration on Ionic buildings, of which the best examples are from the Siphnian Treasury at Delphi (Archaeological Museum, Delphi), constructed shortly before 525 BC. The shallow relief on these works is little more than drawing rendered partly in the round; but the sculptor soon learned how, even in the shallowest relief, to indicate depth by overlapping figures and by bringing details up into the front plane. A dark-painted background helped the illusion; but the effect of the lavish use of colour on flesh, drapery, and backgrounds cannot now be readily appreciated since so little of it has survived in more than ghostly traces.

Hirmer Fotoarchiv, Munchen

Figure 11: The "Kritios Boy," marble kouros, c. 490–480 BC. In the Acropolis Museum, Athens. Height 86 cm.

THE CLASSICAL PERIOD

Early Classical (c. 500–450 BC). This brief period is more than a mere transition from Archaic to Classical; in the figurative arts a distinctive style developed, in some respects representing as much of a contrast with what came afterward as with what went before. Its name—Severe style—is in part an indication that the "prettiness" of Archaic art, with its patterns of drapery and its decisive action, has been replaced by calm and balance. In vase painting and in sculpture, this new tone is evident in the composition of scenes and in details such as drapery, where the fussy pleats of the Archaic chiton give place to the heavy, straight fall of an outer robe called the peplos. The finest artists transformed the verve of the late Archaic

The Severe style

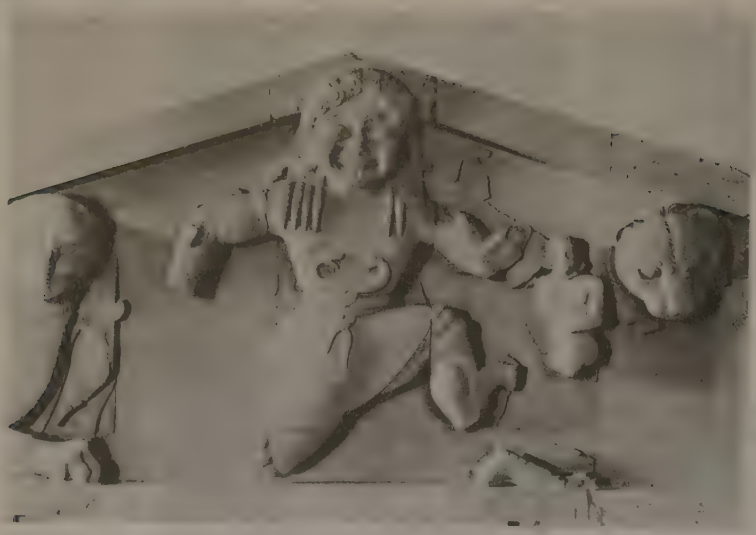


Figure 13: Gorgon from the west pediment of the Temple of Artemis, Corcyra (Corfu), Greece, limestone, *c.* 580 BC. In the Archaeological Museum, Kérkira, Greece.
By courtesy of the Deutsches Archaologisches Institut, Athens

style into more delicate expressions of emotion, and some were clearly checking their work more deliberately against the living model.

The early Classical period saw an impressive series of sculptural works that were excellent in their own right and significant in the continuing development of technical expressive skill and naturalism such as the relief carvings of the so-called Ludovisi Throne (Figure 14). Moreover, for the first time individual artists—and their contributions to technical and stylistic development—can in some cases be positively identified through Roman copies and written descriptions of their works.

Works of the Olympia Master

The finest examples of early Classical architectural sculpture are the works of the Olympia Master, an unidentified artist who decorated the pediments and frieze (Archaeological Museum, Olympia) of the Temple of Zeus at Olympia. In the east pediment, which shows men and women preparing for a chariot race, his figures display the sobriety and calm characteristic of the early Classical period. The men stand in the new, relaxed pose (the weight of the body being carried mainly by one leg) that was to be used by most sculptors throughout the period; and the women wear the peplos, its broad, heavy folds lending severity to the static composition. The west pediment, with a scene of struggling men and centaurs, has something of the rigid formality of the Archaic spirit, but here—and in the metopes that show the labours of Heracles—the artist has acutely observed differences of age in the human bodies and differences of expression—pain, fear, despair, disgust—in the faces (Figure 15). This was something new in Greek sculpture, and, in fact, cannot be readily matched in other works of this period.

The works of Myron

In freestanding sculpture—at this time, more commonly bronze than marble—the works of Myron (of Eleutherae, in Attica), identified through copies, were among the most celebrated of the period. Myron's most famous work is the "Discobolos" (discus thrower), of which a Roman copy (Museo Nazionale Romano) survives. Another of Myron's works surviving in copy is a sculpture of Athena with the satyr Marsyas (Athena in Städtische Galerie, Frankfurt am Main; Marsyas in Lateran Museums, Rome). The interplay of mood and action between the figures in this freestanding group is new, foreshadowed only by the now lost group of the Tyrantslayers erected in Athens at the end of the 6th century. Because bronze was often looted and corrodes easily, the majority of freestanding sculptures from this period have been lost. Some, however, have been rediscovered in the 20th century, the "Poseidon" (National Archaeological Museum, Athens) and the "Charioteer" from Delphi (Archaeological Museum, Delphi), for instance, although they have been eclipsed in fame by the still more remarkable pair of warriors dredged from the

Alinari—Art Resource/EB Inc

Figure 14: "Ludovisi Throne," *c.* 460 BC. In the Museo Nazionale Romano.

Figure 15: "Atlas Brings Heracles the Apples of the Hesperides in the Presence of Athena," marble metope from the east end of the Temple of Zeus at Olympia, *c.* 460 BC. In the Archaeological Museum, Olympia, Greece. 1.60 m × 1.42 m.

By courtesy of the Deutsches Archaologisches Institut—Athens

sea in 1972 and displayed in the Museo Nazionale, Reggio di Calabria. The finer of these latter bronzes (Figure 16), although it probably represents a mortal, has a supernatural glamour and a ferocity quite unlike the calm solemnity conventionally admired in Classical works. This derives partly from the glowing surface of the swelling musculature and the use of inlay for eyes, teeth, and lips.

High Classical period (c. 450–400 BC). Since Roman times, Greek art of the second half of the 5th century BC has been generally regarded as the high point in the development of the Classical tradition. It was the most refined expression of the Greek view of their gods as men and of their men as partaking of the divine. The aesthetic result of this concept was that the bestial or supernatural was abjured in representations of the divine; thus, even a Greek monster, such as the centaur, seems plausible as an image combining humanity and divinity. To some degree, the idealization of human figures was facilitated by the Greeks' traditional concern with proportion and pattern. As a result of the value placed on the ideal image, the representation of extremes (of age or youth, for example, or of deep emotion) and of individuality was ignored or little practiced. Even figures engaged in violent or painful action have a calm, detached expression that modern observers may find chilly and unfeeling. Another reflection of the value placed on the ideal image is an increasing preoccupation with the "heroic nude." From an early phase of Greek art, the artist had shown his interest in man as man rather than as individual. In the Archaic period, the artist studied the visual pattern of the naked male body. When anatomical competence was complete, it was still the abstraction, the pattern, that dictated that his subjects be nude; for it is certain that the average Greek dressed for everyday life and for battle and that only in the exercise ground or the racetrack was the naked body freely revealed.

The idealization of man

During the high Classical period, Athens resumed a position of importance as an artistic centre of the Greek world after years of inactivity. Once most of the Greek homelands were secure from the Persian threat, the funds that had been provided to Athens by the Greek states to lead their defense were turned by the statesman Pericles to the embellishment of Athens itself, and a program of rebuilding temples in the city and countryside was begun. This task attracted sculptors, masons, and other artists to Athens from all over the Greek world. It is largely the work of these artists, under the guidance of Athenian masters, that determined what is now recognized as the high Classical style.

Of the several types of sculpture that flourished during the high Classical period, major statuary is least represented in surviving examples. Phidias, the most influential sculptor of the period, made two huge cult images plated with gold and ivory, the statue of Athena for the Parthenon and a seated statue of Zeus for the temple at Olympia that was one of the seven wonders of the ancient world. These works amazed and overawed viewers through all antiquity, but no adequate copies survive. Another important sculptor of the period, whose work can be seen through copies, was Polyclitus, from Argos. Polyclitus embodied his views on proportion in his "Doryphoros" ("Spear Bearer"), called "The Canon" because of its "correct" proportions of one ideal male form. Unlike freestanding statues, architectural sculpture from the high Classical period has survived in abundance. The Parthenon sculptures must have been executed by many different hands, but, because the overall design was by Phidias, the composition and details undoubtedly reflect his style and instructions. The pedimental figures and frieze, especially, display the Classical qualities of idealization (Figure 17). These allow an approximate assessment of Phidias' style and the importance of his contribution to the establishment of the Classical idiom. About the time that full employment for sculptors in Athens on the Parthenon came to an end, there began a distinguished series of carved relief gravestones for Athenian cemeteries. The general type had been familiar in Archaic Athens, and the practice continued in other parts of Greece through the early Classical period, mainly in the islands and in Boeotia. The new Attic series, with calm and dignified groups of figures in generalized settings of domesticity or

The statues of Phidias and Polyclitus

SCALA/Art Resource, NY

Figure 16: Greek warrior, 5th century BC, one of a pair of bronze statues found in the Mediterranean off Riace, Italy. In the Museo Nazionale, Reggio di Calabria, Italy. Height 2.00 m.

Figure 17: Probably "Leto, Dione, and Aphrodite," marble figures from the east pediment of the Parthenon on the Athenian Acropolis, *c.* 432 BC. In the British Museum, London. Over life-size.

By courtesy of the trustees of the British Museum

leave-taking, exploited effectively the rather impersonal calm in figure and features of the Classical conventions.

The other important class of sculpture, much of which has survived in the original, is the dedicatory—votive reliefs or major works like the "Nike" ("Victory") found at Olympia, made by Paeonius. This work, and others that belong to the last years of the century, such as the frieze from the balustrade of the temple of Athena Nike on the Acropolis at Athens, give a clear indication of progress and change in sculptural style. In the representation of the female body, never before a subject of particular interest to the sculptor (with the distinguished exception of the Olympia Master), true femininity was at last achieved through observation; in these works the figures are no longer like male bodies with the more obvious female characteristics added, which had generally been true of earlier works. Drapery, which had for its patterns been an important element of female figures in the Archaic period, has a heaviness, almost a life of its own in the Parthenon sculptures. By the end of the century, in the Nike balustrade, it is shown pressed tight against the body revealing the forms of the limbs and torso clearly beneath, with brittle, dramatic folds standing clear of the surface. This last style, together with the new approach to the rendering of women's bodies, led quickly to a deliberately sensual effect in statuary and hastened the decline of the unemotional Classical conventions.

Developments in rendering the female body

Late Classical period (c. 400–323 BC). The 4th century saw a dramatic increase of wealth in Greece but less in the hands of the warring states of the 5th century and more concentrated on the periphery of the Greek world—with the western colonies, the eastern Greeks, who continued in close touch with the friendlier Persian provinces, and the increasingly powerful Macedonian kingdom in the north. Macedonian power, culminating in Alexander the Great's annexation of the whole Persian Empire in the third quarter of the 4th century, was to transform Greek art as effectively as it did Greek life and politics. Even before Alexander's accession, however, the seeds of change were sown. The many new centres and patrons for artists may have made it easier for them to break with Classical conventions established in 5th-century Athens or by dominant 5th-century artists like Polyclitus. The trend was toward greater individuality of expression, of emotion, and of identity, leading eventually to true portraiture. The last was encouraged by the ambitions and pride of rulers such as the Macedonian kings or by the royal houses of Hellenized provinces in the western Persian Empire. To the same sources can be traced the new interest in monumental tomb construction. Men were aspiring more openly to divinity, and Greek art was no barrier to its explicit expression. It is clear, however, that artists were conscious of the values that were set in the 5th century, and by no means did they act as revolutionaries in styles or techniques. The development of Greek art was swift but smooth, and personalities lent impetus to the development rather than changing its flow dramatically.

Three names dominate 4th-century sculpture, Praxiteles, Scopas, and Lysippus. Each can be appreciated only through ancient descriptions and copies, but each clearly contributed to the rapid transition in sculpture from Classical idealism to Hellenistic realism. Praxiteles, an Athenian, demonstrated a total command of technique and anatomy in a series of sinuously relaxed figures that, for the first time in Greek sculpture, fully exploited the sensual possibilities of carved marble. His Aphrodite (several copies are known), made for the east Greek town of Cnidus, was totally naked, a novelty in Greek art, and its erotic appeal was famous in antiquity. The "Hermes Carrying the Infant Dionysus" (Archaeological Museum, Olympia) at Olympia, which may be an original from his hand, gives an idea of how effectively a master could make flesh of marble (Figure 18). The reputation of Scopas, from the island of Paros, came from the intensity of expression with which he imbued his figures. Fragments of his work at Tegea (National Archaeological Museum, Athens) show his technique in the deep-sunk eye sockets that characterize his faces and that transform the hitherto passionless features of Classical sculpture into studies of intense emotion. Praxiteles and Scopas seem to typify the new spirit that can readily be discerned in surviving original sculptures. The "Demeter of Cnidus" (British Museum, London; perhaps by the Athenian sculptor Leochares) is Classical in mood, but the features are Praxitelean; and in the reliefs on the Mausoleum (British Museum, London) at Halicarnassus (on which both Scopas and Leochares are said to have worked), the vigour of the battle scenes is heightened by both the intensity of the features and a new, rather flamboyant use of drapery. On Athenian grave reliefs the Classical calm gave place to expressions of controlled but deep emotion. These are styles that can be recognized in places far from Greek soil, as in the relief sarcophagi fashioned by the Greeks for the kings of Sidon in Phoenicia.

The sculpture of Praxiteles

Lysippus, from Sicyon in the northern Peloponnese, was Alexander's favourite sculptor. He was true to the Classical tradition in demonstrating his views on proportion by sculpturing athlete figures in different poses, although his types have heavier bodies and smaller heads than those of the Classical standard set down by Polyclitus. But he adds something to these single figure studies; for the first time they are composed in such a way that the viewer is

The sculpture of Lysippus

Figure 18: "Hermes Carrying the Infant Dionysus," marble statue by Praxiteles, *c.* 350–330 BC (or perhaps a fine Hellenistic copy of his original). In the Archaeological Museum, Olympia, Greece. Height 2.15 m.
Hirmer Fotoarchiv, Munchen

invited to move around them, and they are not tied to a single optimum viewpoint, as even Praxiteles' figures had been. This was an important innovation in the history of sculpture.

Another innovation, in the development of which Lysippus must also have played a vital part, is portraiture; he carved likenesses of Alexander. Nevertheless, portraits of contemporaries were still exceptional, and many early portraits are semi-idealized studies of the great philosophers, statesmen, or poets of the Classical period. And yet, it is clear that by now the use of live models was commonplace, as can be judged from the works or copies that survive and from stories of Praxiteles' use of his mistress Phryne as a model or of Lysippus' brother taking casts from life. By the time of Alexander most of the important problems in the realistic or dramatic treatment of features, pose, and drapery had been solved, leaving to later generations an opportunity only to exaggerate anatomy or expression or to devise sculptural groups of yet greater complexity. Fourth-century sculptors, led by Praxiteles, Scopas, and Lysippus, gathered and expressed the best of what had been learned before of anatomy, pattern, and composition; by adding emotional appeal they can be said to have achieved the logical culmination of the Classical tradition, in which Phidian sculpture in the 5th century was but one brilliant and influential episode.

Hellenistic period. Styles of Hellenistic sculpture were determined by places and schools rather than by great names. Pergamene sculpture is exemplified by the great reliefs from the altar of Zeus (Figure 19), now in East Berlin, and copies of dedicatory statues showing defeated Gauls (Figure 20, bottom). These, like the well-known "Nike of Samothrace" (Figure 21), are masterful displays of vigorous action and emotion—triumph, fury, despair—and the effect is achieved by exaggeration of anatomical detail and features and by a shrewd use of the rendering of hair and drapery to heighten the mood. The "Laocoon" group (Vatican Museums), a famous sculpture of the Trojan priest and his two sons struggling with a huge serpent, probably made by Rhodian artists in the 1st century AD but derived from examples of suffering figures carved in the 1st century BC, is a good example of this applied to a freestanding group (Figure 20, left); and the "Belvedere Torso" (Vatican Museums), much admired in Renaissance Italy, of the effective emphasis of anatomy (Figure 20, right). In vivid contrast, a fully sensual treatment of the female nude was achieved by careful surface working of the marble, and the accentuation of femininity by the incorporation of sloping shoulders, tiny breasts, and high full hips. It is the Hellenistic Aphrodite, such as the "Venus de Milo" (Figure 22), who proliferates in Roman copies. The sculptural groups such as Laocoon were novel, demanding a palatial or sanctuary setting and far removed from earlier two-figure groups or the more nearly comparable but one-view pedimental compositions. The new realism extended to the portrayal of old age, decrepitude,

(Right) By courtesy of the Staatliche Museen zu Berlin, photograph, (left) EDI Studio, Barcelona

Figure 19: *Hellenistic relief sculpture.*
Great altar of Zeus at Pergamum. In the Staatliche Museen zu Berlin. (Left) Reconstruction of the west front. (Right) Frieze detail of Alcyoneus seized by the hair, from the marble relief on the east side of the great altar of Zeus, *c.* 180 BC.

Figure 20: *Vigorous action and dramatic emotion in Hellenistic sculpture.*
(Top left) "Laocoon," marble sculpture by Agesander, Athenodorus, and Polydorus of Rhodes, 1st century AD. In the Vatican Museums. Height 2.41 m. (Top right) "Belvedere Torso," marble by Apollonius, 1st century BC. In the Vatican Museums. (Bottom) "Dying Gaul," marble Roman copy after a bronze original, from Pergamum, c. 230–220 BC. In the Museo Nazionale Romano.
(Top) Alinari—Art Resource/EB Inc., (bottom) Anderson—Alinari from Art Resource/EB Inc

New realism in sculpture

disease, low life, and even the grotesque. Alexandria, in its major and minor (clay) works of sculpture, seems to have been one of the important schools in this genre. For the first time in Greek art, babies were rendered as other than reduced adults. In portraiture, the idealizing tendencies of the 4th century were still strong, and portraits of kings or poets were overlaid by conceptions of kingship or artistry. It was to take Roman patronage to enforce a more brutal realism in portraiture of contemporaries.

Copies and stylistic revivals

Two of the most significant developments in Hellenistic sculpture, however, had nothing to do with the evolution of new styles or types of compositions. The first was the production of accurate copies of earlier works, which began by about 100 BC, in part occasioned by the demand from the Roman West. This production stimulated interest in the styles of the great Classical sculptors and helped to determine the decidedly Classical atmosphere of early imperial art. The second, related development is the creation of original works deliberately in the style of the late Archaic, early Classical, or full Classical periods. This archaizing can be seen as both a reaction against the more exuberant Hellenistic sculptural styles and a response to the new interest in the Classical past.

Greece and Rome. It was Hellenistic art that the great Roman Republic and its early empire came to know and to covet. It was already to some degree familiar to them from the work of the western Greeks in Italy and Sicily, and the Romans formed a closer acquaintance with it in the court of Alexandria and from the profits of their diplomacy and warfare. The flow of works of art and artists to the west began, and the classical styles of early imperial Rome are exactly those of the late Hellenistic Greek world, in many instances executed by the same artists. Thus, in the early empire the majority of known artists' signatures are those of Greeks. The adoption of Greek art by the Roman Empire ensured its continuity in the Western tradition and its eventual transmission, through the Renaissance revival, to the modern world.

(Jo.Bo./Ed.)

Roman and Early Christian

There are many ways in which the term ancient Roman art can be defined, but here, as commonly elsewhere, it is used generally to describe what was produced throughout the part of the world ruled or dominated by Rome until

Figure 21: "Nike of Samothrace," marble statue, *c.* 200 BC. In the Louvre, Paris. Height 2.44 m.
J E Bulloz

around AD 500, including Jewish and Christian work that is similar in style to the pagan work of the same period.

Greek influences

The Romans were always conscious of the superiority of the artistic traditions of their neighbours. Such works of art as were made in or imported into Rome during the periods of the monarchy and the early republic were produced almost certainly by Greek and Hellenized Etruscan artists or by their imitators from the cities of central Latium; and throughout the later republican and the imperial epochs many of the leading artists, architects, and craftsmen had Greek names and were Greek, or at any rate Greek-speaking. References in ancient literature and signatures of artists preserved in inscriptions leave no doubt on this point. According to tradition, the earliest image of a god made in Rome dated from the 6th century BC period of Etruscan domination and was the work of Vulca of Veii. A magnificent terra-cotta statue of Apollo found at Veii (Figure 23) may give some notion of its character. In the 5th, 4th, and 3rd centuries BC, when Etruscan influence on Rome was declining and Rome's dominion was spreading through the Italian peninsula, contacts with Greek art were no longer chiefly mediated via Etruria but, instead, were made directly through Campania and Magna Graecia; paintings and "idealizing" statues of gods and worthies mentioned in literature as executed in the capital during this period were clearly the works of visiting or immigrant Greek artists. The plundering of Syracuse and Tarentum at the end of the 3rd century BC marked the beginning of a flow of Greek art treasures into Rome that continued for several centuries and played a leading role in the aesthetic education of the citizens.

Literature shows that by the middle of the 2nd century BC the Roman forum was thronged with honorific statues of Roman magistrates, which, although none of them has survived, may be assumed to have been carved or cast by Greeks because no native Roman school of sculptors of that time is known. And it is significant that the earliest account of Roman realistic portraits of private individuals is contained in the Greek historian Polybius' description of ancestral *imagines* ("masks") displayed and worn at patrician funerals—a description written about the middle of the 2nd century BC, when the tide of Greek artistic influence was sweeping into Rome and Italy from countries east of the Adriatic, where a highly realistic late-Hellenistic portrait art, which sometimes depicted Roman or Italian subjects, had already blossomed.

Three art forms that expressed the Roman spirit

The first appearance of three art forms that expressed the Roman spirit most eloquently in sculpture can be traced to the Hellenistic Age. These forms are realistic portraiture showing a preference for the ordinary over the heroic or legendary, in which every line, crease, wrinkle, and even blemish was ruthlessly recorded; a continuous style in narrative art of all types; and a three-dimensional rendering of atmosphere, depth, and perspective in relief work and painting. Of these three art forms there is no evidence in the early art of pre-Hellenistic central Italy; and it would be safe to guess that, if Rome had not met them in the homelands of Greek art, it would never have evolved them in its great art of imperial times. But Rome's own contributions to art, if of a different order, were vitally important. Its historical aims and achievements furnished late Hellenistic artists with a new setting and centre, new subjects, new stimuli, a new purpose, and a new dignity. Rome provided the external circumstances that enabled architects, sculptors, painters, and other craftsmen to exploit on a much more extensive scale than before artistic movements initiated in the Hellenistic world, and Rome became a great new patron of art and a great new wellspring of inspiration and ideas.

THE LAST CENTURY OF THE REPUBLIC

Funerary masks

Ancestral *imagines*, or funerary masks, made of wax or terra-cotta, had become extremely individualized and realistic by the middle of the 2nd century BC. The source of this realism is in the impact on Rome of late-Hellenistic iconography; although this use of masks was rooted

J E Bulloz

Figure 22: "Venus de Milo," marble statue of Aphrodite from Melos, *c.* 150 BC. In the Louvre, Paris.

Figure 23: "Veii Apollo," clay statue, *c.* 500 BC. In the Museo Nazionale di Villa Giulia, Rome. Height 1.75 m.
Pellegrini—Crimoldi

in ancient Roman social and religious practice, there is no basis for a belief that the Romans and Etruscans had, from early times, been in the habit of producing death masks proper, cast directly from the features of the dead. It was undoubtedly their funerary customs that predisposed the Romans to a taste for portraits; but it was not until around 100 BC that realistic portraiture, as an art in its own right, appeared in Rome as a sudden flowering, and to that time belong the beginnings of the highly realistic heads, busts, and statues of contemporary Romans—in marble, stone, or bronze—that have actually survived. Coin portraits of public personages, whose names and dates are recorded, greatly assist in determining a chronological sequence of the large-scale likenesses, the earliest of which can be attributed to the period of Sulla (82–79 BC). The style reached its climax in a stark, dry, linear iconographic manner that prevailed around 75–65 BC and that expressed to perfection current notions of traditional Roman virtues; of this manner, a marble head of an elderly veiled man in the Vatican is an outstanding illustration (Figure 24, top left). Shortly thereafter, an admiration for earlier phases of Greek art came into fashion in the West, and verism was toned down at the higher social levels by a revival of mid-Hellenistic pathos and even by a classicizing trend that was to stamp itself upon Augustan portraits. Meantime, in sepulchral custom, the ancestral bust had become an alternative to the ancestral mask, a development exemplified in a marble statue of a man wearing a toga and carrying two such busts in the Capitoline Museums at Rome (Figure 24, bottom); and portrait busts and figures carved on numerous stone and marble grave stelae (slabs or pillars used for commemorative purposes), characteristic of the late republican epoch, suggest the persistence of a preference for severe pose in middle-class and humbler circles. Furthermore, there are some 1st-century-BC portraits that suggest that the making of death masks proper (arguably a sophisticated idea) was occasionally practiced at this time (Figure 24, top right). None of the vivid Etruscan portraits, such as a bronze orator popularly called the "Arringatore" (Museo Archeologico) at Florence (Figure 25) and a terra-cotta married pair on the lid of a cinerary chest (for ashes of the dead) in the Museo Etrusco Guarnacci, at Volterra, is earlier than *c.* 100 BC; works of that type may be reckoned as provincial imitations of the new metropolitan, 1st-century-BC portrait style.

Narrative reliefs

There are no narrative reliefs from Rome that can confidently be assigned to a date before 100 BC. The only definitely dated 2nd-century-BC relief depicting an episode from contemporary Roman history, a frieze with the Battle of Pydna on Lucius Aemilius Paulus' victory monument at Delphi, was worked in 168 BC in Greece. The most familiar republican example of this form of art as practiced in the West is frieze decoration (partly in the Louvre, and

By courtesy of (top right) the Deutsches Archaologisches Institut Rom; photographs, (top left) Anderson—Alinari from Art Resource/EB Inc., (bottom) Alinari—Art Resource/EB Inc

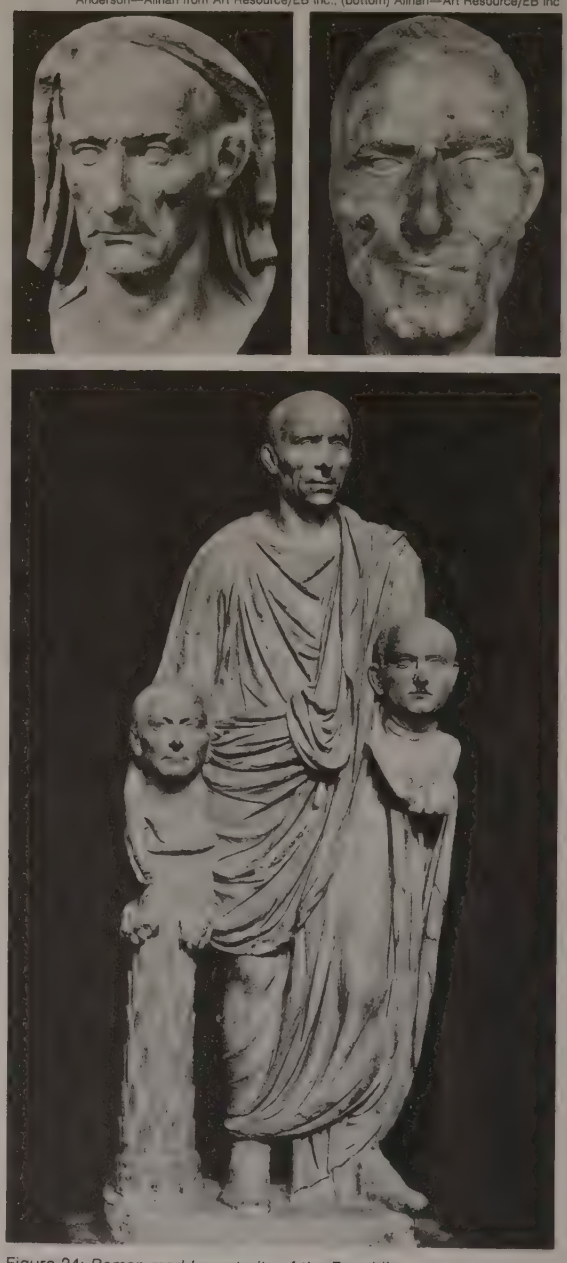


Figure 24: *Roman marble portraits of the Republic.* (Top left) Head of an elderly veiled man, c. 75–65 BC. In the Vatican Museums, Rome. Life-size. (Top right) Portrait in death-mask style, 1st century BC. In the Museo di Antichità, Turin, Italy. Life-size. (Bottom) A Roman patrician with the busts of his ancestors, 1st century AD. In the Capitoline Museums, Rome. Life-size.

Figure 25: Bronze statue of an orator ("Arringatore"), *c.* 150 BC. In the Museo Archeologico, Florence. Height 1.80 m.
Anderson—Alinari from Art Resource/EB Inc

partly in the Glyptothek at Munich) from the so-called Altar of Ahenobarbus, which has been shown to have no sure connection either with an altar or with any of the Ahenobarbi. In these, prosaic documentation of Roman census procedure is juxtaposed with depictions of Greek sea nymphs, a conjunction of literalism and borrowed poetry typical of subsequent Roman art.

Funerary narrative sculpture

Funerary narrative sculpture of the late republic is exemplified in a monument of the Julii, at Saint-Rémy (Glanum), France. The base of this structure carries four great reliefs with battle and hunt scenes that allude not only to the mundane prowess of the family but also to the otherworldly victory of the souls of the departed over death and evil, since figures of the deceased, accompanied by personifications of death and victory, merge into one of the battle scenes. It is possible that these highly pictorial reliefs were partly based on lost Hellenistic monumental paintings, for southern Gaul had direct connections with Greek lands east of the Adriatic.

THE EMPIRE

Augustan Age. The hallmark of portraits of Augustus is a naturalistic classicism. The rendering of his features and the forking of his hair above the brow is individual. But the Emperor is consistently idealized and never shown as elderly or aging. A marble statue from Livia's Villa at Prima Porta (in the Vatican), which presents him as addressing, as it were, the whole empire, is the work of a fine Greek artist who, while adopting the pose and proportions of a classical Hellenic statue, perfectly understood how to adopt these to the image that Augustus cultivated as emperor (Figure 26). On his ornate cuirass (armour protecting the chest and back), Augustus' aims and achievements are recorded symbolically in a series of figure groups. A marble portrait statue found on the Via Labicana (Museo Nazionale Romano) represents the Emperor as heavily draped and veiled during the act of sacrificing as *pontifex maximus* ("chief priest"); and a bronze head from Meroe in The Sudan (British Museum), the work of a Greco-Egyptian portraitist, depicts him as a Hellenistic king. Of the female portraits of the period, one of the most charming is a green basalt head (Louvre) of the Emperor's sister, Octavia, with the hair dressed in a puff above the brow and gathered into a bun behind—a popular coiffure in early Augustan times. In many respects, the noblest of all Roman public monuments that were adorned with sculpture is the Ara Pacis Augustae ("Augustus' Altar of Peace"), founded in 13 BC and dedicated four years later (Figure 27). It stood in the Campus Martius and has been restored, with different orientation, not far from its original site. On its reliefs—significantly of Luna marble, a white marble quarried in Italy and not, as had earlier been the case, imported from Greece—it set a standard of distinction surpassed by no later work, with the harmonious blending of contemporary history, legend, and personification, of figure scenes and decorative floral motifs. The altar proper was contained within a walled enclosure, measuring about 38 by 34 feet (11½ by 10½ metres), with entrances on east and west. On the upper part of the external faces of the south and north precinct walls ran a frieze representing the actual procession (of Augustus, members of his family, officers, priests, magistrates, and the Roman people) to the altar's chosen site on its foundation day (July 4, 13 BC), when sacrifice was offered in thanksgiving for the Emperor's recent return to Rome from the provinces. On either side of the western entrance was a depiction of Augustus' prototype Aeneas sacrificing on his homecoming to the promised land of Italy, and, since Augustus was also hailed as Rome's second founder, a depiction of the suckling of the twins, Romulus and Remus, by the she-wolf. The eastern entrance was flanked by personifications of Roma and of Mother Earth with children on her knees flanked by figures symbolizing air and water (Figure 27, bottom). On the exterior of the

The Ara Pacis Augustae

Alinari—Art Resource/EB Inc

Figure 26: Augustus of Prima Porta, marble statue, *c.* 20 BC
In the Vatican Museums, Rome. Height 2.03 m.

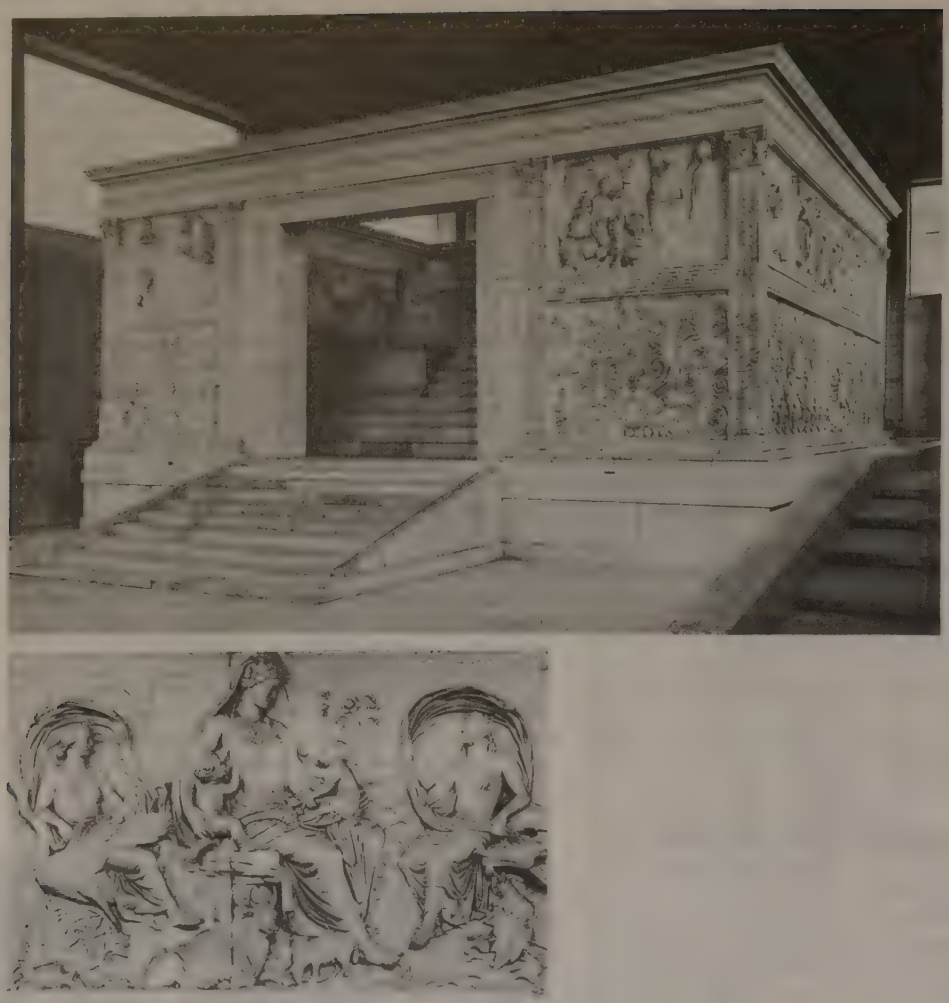


Figure 27: Ara Pacis, Rome, 13 BC. (Top) View of whole altar. (Bottom) "Mother Earth with Air and Water," marble relief on the east exterior wall of the Ara Pacis. Height 1.57 m.
(Top) Alinari—Art Resource/EB Inc . (bottom) Fototeca Unione, Rome

walls, beneath all these figure scenes, was a magnificent dado filled with a naturalistic pattern of acanthus, vine, and ivy, perhaps a translation into marble of a gorgeous carpet or tapestry used in the ceremony. Swags of fruit and flowers that decked the interior faces of the precinct walls may represent real swags that were hung on the temporary wooden altar erected for the foundation sacrifice. The procession was continued in a much smaller frieze on the inner altar, from which figures of Vestal Virgins and of sacrificial victims and their attendants have been preserved. Delightful studies of imperial and other children and such homely incidents as conversations between persons taking part in the procession introduce an element of intimacy, informality, and even humour into this solemn act of public worship. The Ara Pacis, in fact, sums up all that was best in the new Augustan order—peace, serenity, dignity without pompousness, moderation and absence of ostentation, love of children, and delight in nature. The style of the altar's floral decoration strongly suggests that the sculptors who carved it were Greeks from Pergamum.

Julio-Claudian period. The imperial portraiture of Tiberius and Caligula was generally precise but academic work, but some of the female court portraits reflect not only the fashions for elegant simplicity and extreme elaboration in female coiffure but also a subtle poetry. Two possible extremes of tone are clearly marked by the contrasting busts of Claudius and Nero, the former uncomfortably uncompromising, the latter flatteringly Hellenic. In the relatively few public monuments dating from this period to include sculpture, none reveals any novel development.

Flavian period. In the emperor Vespasian's portraits, something of the old, dry style returned. This can be observed in his striking likeness on one of two historical reliefs (Vatican Museums) that were unearthed in Rome near the Palazzo della Cancelleria. A similarly sketchy and impressionistic handling of the hair is found on the emperor Titus' portraits, whereas the third Flavian emperor, Domitian, affected a more pictorial hairdo in imitation of the coiffure introduced by Nero. Still more picturesque are the female hair styles of the time, which display piles of corkscrew ringlets or tight, round curls (Figure 28). The Cancelleria reliefs date from the close of Domitian's reign and depict, respectively, Vespasian's triumphal entry and reception in Rome in AD 70 and Domitian's *profectio* ("setting out"), under the aegis of Mars, Minerva, and Virtus, for one of his northern wars. They are worked in a two-dimensional, academic, classicizing style that is in marked contrast with the vivid, three-dimensional rendering of space and depth, with brilliant interplay of light and shade, on the panels of the Arch of Titus in the Roman Forum. The latter reliefs, which present two excerpts from Titus' triumph in Palestine, were carved in the early 80s. The late Domitianic classicizing manner appears again in the frieze of the Forum Transitorium, which the emperor Nerva completed. This conflict of relief styles within the Flavian period is but one illustration of the ceaseless, unpredictable ebb and flow of different aesthetic principles throughout the history of imperial art.

Age of Trajan. In portraits of Trajan, a deepening of the bust, which was already seen in the later Flavian period, was carried a stage further; there is a new fluidity in the molding of the face; in the hair, which is plastered down across the brow, there is a partial revival of the late republican linear style. Aesthetically, one of the finest known likenesses of the Emperor is a marble head from

Figure 28: Portrait of a woman of the Flavian period, marble, c. AD 90. In the Capitoline Museums, Rome. Life-size.
Cesare Faraglia

Ostia (Ostia Museum). On his monumental column there is a series of less idealized and probably more faithful renderings of his features. The coiffures of Trajanic ladies are, if anything, even more elaborate and extravagant than those of their Flavian predecessors.

Reliefs of Trajan's Column

The reliefs of Trajan's Column, illustrating the two Dacian campaigns of 101–102 and 105–106 and winding up the shaft in a spiral band of Parian marble three feet (one metre) wide, are generally recognized to be the classic example of the continuous method of narration in Roman art. The episodes merge into one another without any punctuation, apart from an occasional tree; Trajan appears again and again in different situations, activities, and costumes. A statuesque figure of Victory separates the histories of the two wars. There are 23 spirals and about 2,500 figures. A high level of technical accomplishment is maintained throughout, and the interest and excitement of the theme never flag. Since the figures of men and animals had to be distinguished from a distance, they are inevitably overlarge in proportion to their landscape and architectural settings; and in order to avoid awkward empty spaces along the upper edges of the band and to preserve an allover, even, tapestry-like effect, background figures in the scenes are reared in bird's-eye-view perspective above the heads of those in the foreground. These carvings must be visualized as once brightly painted, with weapons and horse trappings added in metal. The sources of the scenes were possibly wartime sketches made by army draftsmen at the front, but the fusing together of those isolated pictures into a single scroll was the work of a single master artist, perhaps Apollodorus of Damascus, who designed the whole complex of Trajan's forum, basilica, and column.

The column (the interior of which contains a spiral staircase; Figure 29) had first been intended primarily as a lookout post for viewing Trajan's architectural achievements—his forum and its adjacent markets, to accommodate which he sliced away the slope of the Quirinal Hill. By the time of its dedication in 113, when the relief bands had been added and an eagle planned for the top of the capital, it had become a war memorial. Finally, it became Trajan's future tomb, crowned by his statue (which was later replaced by that of St. Peter) and containing a funerary chamber for the urns holding his and his consort's ashes.

Arch of Constantine

To the last years of Trajan's reign or to the early years of that of his successor should be attributed four horizontal panels that adorn the main passageway and the attic ends of the Arch of Constantine in Rome. If fitted together they would form a continuous frieze of three main scenes, which are, from left to right, an imperial triumphal entry, a battle, and the presentation to the Emperor of prisoners and the severed heads of captives by Roman soldiers. It seems clear that these sculptures were made between around 115 and 120, perhaps for the Temple of Divus Trajanus and Diva Plotina that was erected by Hadrian just to the north of the column. The presence on this frieze of chain-mail corselets, rarely seen on Trajan's Column, seems to indicate that that type of armour, so common under the Antonines (see below *Antonine and Severan periods*), first came into general use in late Trajanic or early Hadrianic times. These reliefs do not depict realistic fighting, as do those of the column, but a kind of ideal or dramatized warfare, with the Emperor himself participating in the melee and the soldiers wearing plumed and richly embossed parade helmets; the scenes melt into one another with total disregard of spatial and temporal logic.

Trajan's arch at Beneventum

A third example of Trajanic monumental sculpture is the relief decoration of the Arch of Trajan at Beneventum (Benevento), which is covered with pictorial slabs, the subjects of which are arranged to carry out a carefully balanced and nicely calculated order of ideas. Those on the side facing the city and on one wall of the passageway present themes from Trajan's policy and work for Rome and Italy; those on the side toward the country and on the other wall of the passageway allude to his achievements abroad. With two exceptions, where a pair of scenes forms a single picture, each panel is a self-contained unit. The reliefs already show something of the classicizing, two-dimensional character of Hadrianic work. Indeed, it seems likely that, although the arch itself was either decreed or dedicated in 114 or 115, some of the panels in which Hadrian is given a peculiar prominence were not carved until the early years of the latter's principate.

The frieze of a great, circular Tropaeum Trajani, set up in the Dobruja (Romania) to commemorate victories over the Dacians, contains a series of metopes (a decoration in a Doric frieze) carved with figure scenes in a naïve, flat, linear style that suggests the hands of army artists of provincial origin.

Age of Hadrian. In the iconography of the age of Hadrian, certain Hellenizing features—the wearing of a

Anderson—Alinari from Art Resource/EB Inc

Figure 29: Trajan's Column, memorial with marble reliefs illustrating the two Dacian wars of 101–102 and 105–106; AD 106–113. In Trajan's Forum, Rome.

short Greek beard by the males and the adoption by the females of a simple, classicizing coiffure—are harmonized with new experiments. The depth of the bust increases, there is greater plasticity in the modelling of the face, the men's curly hair and beards are pictorially treated, and the irises and pupils of the eyes are marked in. Many marble portraits of the Emperor survive from all over the empire, but of his likenesses in bronze only one is extant—a colossal head recovered from the Thames River in London (British Museum), torn from a statue erected in the Roman city and probably the work of a good Gaulish sculptor. Portrait statues of Hadrian's Bithynian favourite, Antinoüs, reveal a conscious return in the pose and proportions of the body to Classical Greek standards, combined with a new emotionalism and sensuousness in the rendering of the head.

Monumental reliefs of the age of Hadrian

The monumental reliefs of Hadrian's day cannot vie with those of his predecessors. The most interesting and perhaps the earliest of them are two horizontal slabs once exposed in the Roman Forum but later transported to the shelter of the Curia. Both carry on one side similar figures of victims for the Suovetaurilia sacrifice and on the other side different historical scenes: in the one case, Hadrian doling out the *alimenta* ("poor relief") to Roman citizens, in the presence of a statuary group of Trajan and Italia with children; in the other case, the burning of debt registers. At one end of each of these scenes is carved a figure, on a base, of the legendary Greek musician Marsyas, whose statue in the Forum may once have been in part enclosed by the panels. In the background of both historical pictures are carved in low relief various buildings in the Roman Forum that can be identified. The two scenes display the characteristically Hadrianic two-dimensional style, as do three large panels (Palazzo dei Conservatori, Rome), with the Emperor's head restored and depicting an imperial triumphal entry, an *adlocutio,* and an apotheosis, respectively—somewhat rigid, academic works. Eight medallions gracing the Arch of Constantine give pleasantly composed and lively, if Hellenizing, pictures of sacrifice and hunting (Figure 30). Some of them depict Antinoüs accompanying the Emperor, whose portraits have been recut as likenesses of Constantine the Great and of his colleague Licinius. Finally, historical reliefs found at Ephesus (now in the Neue Hofburg, Vienna)—one of the very few examples of provincial state reliefs that have survived—may be claimed as late Hadrianic (not as of the period of Marcus Aurelius, to which many critics have assigned them).

In Rome and Italy during the second quarter of the 2nd century, interment began to supersede cremation as a method of disposing of the dead, and Hadrian's reign saw the beginnings of a long line of carved sarcophagi that constituted the most significant class of minor sculptures down to the close of the ancient Greco-Roman world.

Antonine and Severan periods. Portraits of Antonine imperial persons, of which a bronze equestrian figure of Marcus Aurelius on the Capitol (Figure 31) and a great marble bust of Commodus as Hercules in the Palazzo dei Conservatori are perhaps the most arresting examples, display a treatment of hair and beard, deeply undercut and drilled, that grew ever more pictorial and baroque as the 2nd century advanced. This produced an impression of nervous restlessness that contrasts with the still, satin smoothness of the facial surfaces, particularly in the iconography of Commodus. To all this picturesqueness, Septimius Severus added yet another ornamental touch—the dangling, corkscrew forelocks of his patron deity, Sarapis. The female hairstyles of the time are characterized first by a coronal of plaits on top (Faustina the Elder), next by rippling side waves and a small, neat bun at the nape of the neck (Faustina the Younger, Lucilla), and then by stiff, artificial, "permanent" waving at the sides and a flat, spreading "pad" of hair behind (Crispina, Julia Domna).

Figure 31: Bronze equestrian statue of Marcus Aurelius, in the Piazza del Campidoglio, Rome, *c.* AD 173. Height 5.03 m.
Alinari—Art Resource/EB Inc

State reliefs

Of the state reliefs of this epoch, the earliest are on the base (in the Vatican) of a lost column set up in honour of Antoninus Pius and Faustina the Elder. The front bears a dignified, classicizing scene of apotheosis: a powerfully built winged figure lifts the Emperor and Empress aloft, while two personifications, Roma and Campus Martius, witness their departure. On each side is a *decursio,* or military parade, in which the riders farthest from the spectator appear not behind the foot soldiers but high above their heads—a remarkable instance of the bird's-eye-view per-

Alinari—Art Resource/EB Inc

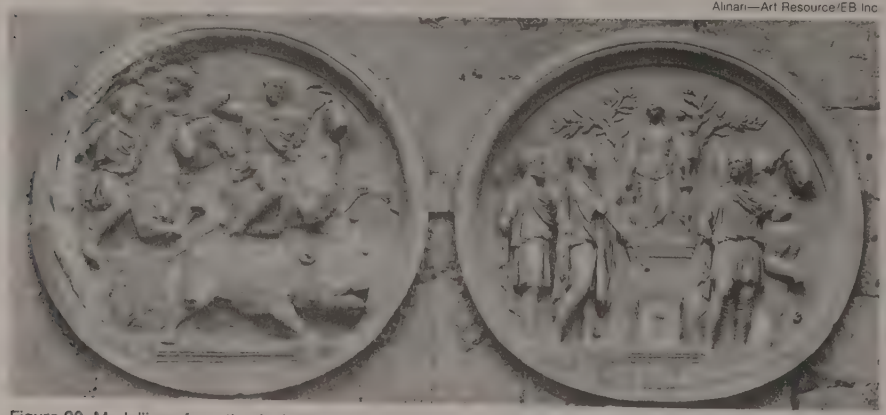


Figure 30: Medallions from the Arch of Constantine, Rome; medallions date from AD 117–138.

spective carried to its logical conclusions. All the figures in these side scenes are disposed on projecting ledges, a device employed again about 20 years later on Marcus Aurelius' Column. Eleven rectangular sculptured panels—similar to those on the Arch of Trajan at Beneventum but displaying greater crowding of figures, livelier movement, and a pronounced effect of atmosphere and depth—depict official occasions and ceremonies in the career of Marcus. Three of these are in the Palazzo dei Conservatori, Rome (Figure 32); the other eight are on the attics (low stories or walls above the cornice of the facade) of the Arch of Constantine. These two sets of panels represent two separate series and may have been carved for two (now lost) distinct triumphal arches. The contrast in style between the spiral reliefs of Marcus Aurelius' Column, put up under Commodus and depicting Marcus Aurelius' northern campaigns, with those of its Trajanic predecessor, is a measure of the change of mood that the Roman world experienced during the course of the 2nd century. The diminished proportions of the squat, doll-like figures, their herding together in closely packed, undifferentiated masses, their angular, agitated gestures, and the stress laid throughout on the horror and tragedy of war suggest that the empire is facing an unknown future with diminished security and that man is at the mercy of some unaccountable power, the supreme embodiment of which is an awe-inspiring winged, dripping figure, personifying the rainstorm that saved the Roman army from perishing from thirst. Again, in the imperial *adlocutiones* that punctuate this frieze, where the Emperor stands in a strictly frontal pose high above the heads of his audiences, can be seen a remarkable return (but probably not a conscious return) to the conventions employed in primitive art for expressing the concept of the ruler as transcendental being.

The spirit of the times is reflected no less vividly in carved sarcophagi. Their themes—familiar myths, battles, hunts, marriages, and so on—allude allegorically to death and the destiny of the soul thereafter. The classicizing, statuesque tradition is also maintained in late 2nd- and early 3rd-century columned sarcophagi, originating in the workshops of Asia Minor but freely imported into, and sometimes imitated in, Rome and Italy. On such pieces single figures or small groups of figures occupy niches between colonnettes. Among the most impressive examples is a great sarcophagus at Melfi, in Puglia, Italy, with a couch-shaped lid, on which the figure of a girl lies prostrate in the sleep of death.

Mansell—Alinari from Art Resource/EB Inc

Figure 32: Marcus Aurelius in a quadriga (four-horsed chariot) entering Rome in triumph, from a marble relief in the Palazzo dei Conservatori, Rome, *c.* AD 176. Height 3.23 m.

The novel features that have been noted in the reliefs of Marcus Aurelius' Column were worked out more completely in those of the official monuments set up to honour Septimius Severus, both in Rome and abroad. In the arch erected in 203 at the northern end of the Roman Forum are found crowded masses of small figures in broad bands of relief, perhaps reflecting a style of documentary painting; in the smaller Porta Argentariorum in Rome, erected by bankers and cattle dealers in honour of the Emperor in the following year, there are stiff, hieratic, funeral poses; and above all in the still more remarkable four-way arch set up at Leptis (Lepcis) Magna in Tripolitania to commemorate a visit of about 203 is a pier decorated with a stylized bird's-eye view of an Oriental city under siege and (also on the piers) weirdly elongated representations of captives. The deeply undercut and drilled vine-scroll ornament here and in the Severan basilica nearby is similar to that found in Asia Minor, whence sculptors had doubtless been imported.

The tension between naturalism and schematization

3rd and 4th centuries. A new tension between naturalism and schematization marks the history of late-antique portraiture. In likenesses of Alexander Severus, the facial planes are simplified, and the tumbling curls of the 2nd-century baroque have been banished in favour of a skullcap treatment of the hair and sheathlike rendering of the beard. Toward the middle of the 3rd century, under Philip the Arabian and Decius, this clipped technique in hair and beard was combined with a return to something of the old, ruthless realism in the depiction of facial furrows, creases, and wrinkles. For a time, Gallienus reinstated the baroque curls and emotional expression, but in the later decades of the century the schematic handling of hair, beards, and features reappeared. Finally, in the clean-shaven faces of Constantine the Great and his successors of the 4th and early 5th centuries, the conception of a portrait as an architectonic structure came to stay; and the naturalistic, representational art of the Greco-Roman world was exchanged for a hieratic, transcendental style that was the hallmark of Byzantine and medieval iconography (Figure 33). The hair is combed forward on the brow in rigid, striated locks, and the eyes are unnaturally enlarged and isolated from the other features. The face is so formalized that the identification of any given portrait becomes a problem. A colossal bronze emperor (near the church of S. Sepolcro, Barletta), for example, has been given the names of several different rulers of the late 4th and early 5th centuries. Throughout these centuries the favourite female coiffure shows a plait or twisted coil of hair carried across the back and top of the head from neck to crown, while under Constantine there was a brief revival of the two Faustinas' styles.

Throughout the 3rd and 4th centuries, carved sarcophagi carry on the story of relief work. Aesthetically, the most notable 3rd-century example is an allover tapestry-like battle piece (Ludovisi Collection, Museo Nazionale Romano, Rome), which possibly was made for Decius' son Hostilian.

Of 3rd-century state reliefs in Rome, virtually nothing has survived. Narrow historical friezes carved for the Arch of Constantine, completed for the celebrations of his *decennalia* (10th anniversary of his reign) in 315, show dwarfish, dumpy, niggling figures. Both these reliefs and those of the slightly earlier Arch of Galerius at Thessalonica look as though they had been worked by artists whose experience had been confined to the production of small-scale sculptures. The last examples of Roman carving are reliefs on the base of an obelisk of Theodosius in the Hippodrome at Constantinople, where the Emperor and members of his court, ranged in rigid, hieratic poses, watch the shows. Few original portions are extant of the spiral relief bands that entwined columns of Theodosius and Arcadius in Constantinople.

Minor forms of sculpture. Of the minor forms of sculpture, none is more attractive than the art of modelling—

Figure 33: Marble colossal head of Constantine the Great, part of the remains of a giant statue from the Basilica of Constantine (formerly the Basilica of Maxentius) in the Roman Forum, Rome, *c.* AD 313. In the Capitoline Museums, Rome. Height of the head 2.41 m.

Hirmer Fotoarchiv, Munchen

Stucco work

in relief or in the round—in fine, white stucco. Decorative stucco work was cheaper and easier to produce than carving in stone or marble. Soft and delicate in texture, it was equally elegant whether left white or gaily painted. In domestic architecture it was a useful alternative or accessory to painting; notable are such examples as a pure white, exquisite vault decoration showing ritual scenes with small-scale figures, from a late republican or early imperial house near the Villa Farnesina in Trastevere (Museo Nazionale Romano); handsome pairs of large white griffins, framed in acanthus scrolls against a vivid red ground, in the late republican House of the Griffins on the Palatine; and a frieze depicting the story of the *Iliad*, in white figures on a bright blue background, in the House of the Cryptoporticus, or Homeric house, at Pompeii. For the use of this technique in palaces, the figure work in Domitian's villa at Castel Gandolfo in the Alban hills can be cited; it can be found in such public buildings as the Stabian and Forum baths and the Temple of Isis at Pompeii. The loveliest and most extensive stucco relief work in a semiprivate shrine is that in the underground basilica near the Porta Maggiore, Rome, where the scenes all allude to the world beyond the grave, to the soul's journey to it, and to the soul's preparation for it in this life (Figure 34). Some of the best surviving stuccos are in tombs: the tomb of the Innocentii and the tomb of the Axe under the church of S. Sebastiano on the Via Appia; the tombs of the Valerii and the Pancratii on the Via Latina (in the latter, stucco work is attractively combined with painting in the flat); and the tomb of the Valerii under St. Peter's, Rome, where the interior walls of both the main and subsidiary chambers are almost completely covered with recesses, niches, and lunettes (semicircular or crescent-shaped spaces) containing stucco figures. The Vatican tomb of the Valerii must be reckoned as a classic place for the study of this delightful and all too scantily represented branch of Roman art.

Ivory and wood carving

Ivory was another popular material for minor sculpture. It was worked in the round, in relief, and in such forms as small portraits, figurines, caskets, and furniture ornaments, of which the carved plaques composing the throne, or "Cathedra of Maximianus," at Ravenna (probably 6th century) provide a notable instance. The consular and other diptychs comprise one of the most distinctive types of ivory relief work in the 4th and 5th centuries. Among them are masterpieces that kept alive the traditions of Hellenistic carving, such as a diptych of the Symmachi and Nicomachi families (one leaf of which is in the Victoria and Albert Museum, London, and the other in the Musée de Cluny, Paris), and some outstandingly fine examples of late antique portraiture, such as the Probus diptych at Aosta (Cathedral treasury) with a double portrait of Honorius, the Felix diptych in Paris (dated 428), and one of Boethius, consul in 487, at Brescia (Civico Museo dell'Età Cristiana). Fine examples of wood carving are panels with biblical scenes on the 5th-century door of the church of Sta. Sabina on the Aventine.

Gem carving

Many types of carving in precious stones were practiced by Roman-age craftsmen, and it is to them that the credit goes for the majority of intaglios that have survived from ancient times. (Intaglios are engraved or incised figures depressed below the surface of the stone so that an impression from the design yields an image in relief.) The widespread taste for them is reflected in the many existing glass-paste imitations reproducing their subjects, which include portraits of both imperial and private persons, and a large variety of divine and mythological groups and figures, personifications, animals, etc. Many bear the signatures of Greek artists.

The most impressive series of Roman gems consists of cameos representing imperial persons. These are miniature reliefs cut in precious stones with different coloured strata (so that the relief is of a different colour from the ground), whereas intaglios, like the ancient seals mentioned earlier, were reliefs, as it were, in reverse, cut into the surface so that a true relief only emerges from an impression. Among the earliest surviving examples of the great imperial cameos are the Blacas onyx (British Museum, London), portraying Augustus in the guise of Jupiter; the Gemma Augustea, a sardonyx (an onyx with parallel layers of sard) in the Kunsthistorisches Museum, Vienna, and the Grand Camée de France, a sardonyx in the Bibliothèque Nationale, Paris, which were probably carved under Caligula and present, respectively, the apotheosis of Augustus and of Tiberius, the latter with Divus Augustus, also; and a sardonyx cameo of Claudius with Jupiter's aegis (Royal Art Collection at Windsor Castle). Late antique examples of the craft are a rectangular sardonyx (city library at Trier), portraying Constantine the Great and members of his house and an onyx with busts of Honorius and Maria (Rothschild Collection, Paris).

Other varieties of carving in precious stones are represented by a miniature head of a girl (British Museum) wearing the hair style of Messalina and Agrippina the Younger, which is cut in *plasma;* an onyx vase in the Braunschweigisches Landesmuseum, Braunschweig, possi-

Alinari—Art Resource/EB Inc.

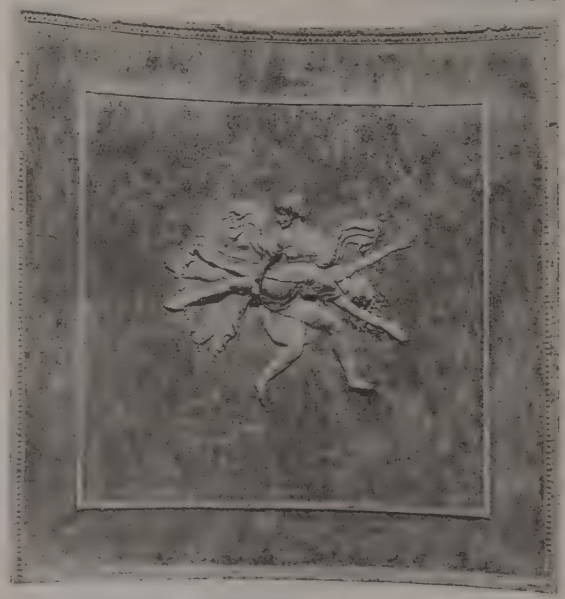


Figure 34: Detail of stucco decoration in a vaulted chamber of a subterranean basilica near Porta Maggiore, Rome, mid-1st century AD.

bly of the 1st century, depicting an emperor and empress as Triptolemus and Demeter; and a late-antique vase, carved in honey-coloured agate and decorated front and back with a naturalistic vine and with the head of Pan, cupped in acanthus, on either shoulder (Walters Art Gallery, Baltimore).

"Cameo glass"

Closely akin to cameos and vessels cut in precious stones are their substitutes in opaque "cameo glass," worked in two layers, with the designs standing out in white against a dark-blue or bright-blue background. To this class belong a blue vase from Pompeii (Museo Archeologico Nazionale), with Cupids gathering grapes; the Auldjo Vase (British Museum, London), with an exquisitely naturalistic vine; and the celebrated Portland Vase, also in the British Museum, the scenes on which have always been the subject of scholarly controversy but are generally supposed to depict myths relating to the afterlife. Similar imitations of carving in precious stones are late antique *diatreta* ("cage cups"), the decoration of which is cut back from the outer surface of the mold-cast blank. This openwork ornamentation sometimes represents the crisscross meshes of a net, while on other vessels it consists of an elaborate figure scene, the design in either case being very deeply undercut and, for the most part, only connected with the background by short shanks of glass. Of the figured examples, the most spectacular surviving specimens are a dark-blue *situla* ("bucket") with a hunting scene (treasury of St. Mark's, Venice) and a dull-green cup presenting the story of Lycurgus ("Rothschild Vase," British Museum). Of the other types of glass with figured decoration, molded cups with gladiatorial and circus scenes are characteristic of the early-imperial period; and the 4th-century glassworker's craft is represented by vessels with cut or incised designs. Among the most important centres of glass production under the empire were Syria, Alexandria, and the Cologne region.

Figured pottery

Figured terra-cotta tablewares (terra sigillata—a term often incorrectly stretched to cover plain wares) were cheaper versions of costly decorated silverwares. During the last century of the republic and in the early decades of the 1st century of the empire, Arretium (Arezzo) was the most flourishing centre of the manufacture of a fine type of red-gloss pottery (Figure 35). As signatures on the pots

By courtesy of the trustees of the British Museum

Figure 35: Arretine ware bowl with a design of the seasons in relief. Made in the factory of Gnaeus Ateius, *c.* 10 BC. In the British Museum. Height 17.8 cm.

reveal, Italian firms often employed Greek and Oriental craftsmen, and the mythological and floral themes of the vessels' molded ornamentation owe much to the inspiration of Hellenistic art.

Gaulish pottery

From shortly before the mid-1st century AD onward, the markets enjoyed by Italian fabrics were captured by the products of potteries now established in southern, central, and eastern Gaul. These manufactured cheaper, more mass-produced, and aesthetically inferior red-gloss and black-gloss wares, popularly known as "Samian," some varieties of which continued into the 4th century. The decoration of Gaulish pots was, for the most part, molded; but some vessels carry applied motifs made in individual molds, and others show designs incised to counterfeit cut glass. Yet another type of ornament was carried out in the barbotine technique, by which relief work was produced by trailing liquid clay across the surface of the pot. As regards the content of the decoration, themes from daily life were added to traditional subjects based on Greco-Roman mythology and on natural history. The *E* barbotine hunt cups (produced mainly at Castor, Northamptonshire) are the highlight of the native Romano-British potter's craft.

A late-antique class of red-gloss pottery, known as late A ware, with scenes in relief from Greek mythology and from Roman spectacles, was manufactured in a southern Mediterranean area, probably Egypt. (J.M.C.T./Ed.)

EARLY CHRISTIAN

Early in the 20th century it was thought that Christian art began after the death of Christ or, at least, in the second half of the 1st century AD. But later discoveries and studies showed that a truly Christian art—that is, with a style quite distinctive from Pagan Roman art—did not exist before the end of the 2nd or beginning of the 3rd century. When it ended, or rather developed into something else, is harder to say. Early Christian art penetrated all the provinces of the Roman Empire, adapting itself to existing pagan art. It subsequently created its own forms, which varied according to local stylistic evolution. The new capital at Constantinople (ancient Byzantium), founded by the emperor Constantine the Great (306–337), was to be an important centre of art. The art produced there, now known as Byzantine art, extended throughout the entire Christian East. It is customary to distinguish early Christian art of the West or Latin part from the Christian arts of regions dominated by the Greek language and to consider the latter as proto-Byzantine, while acknowledging, however, a certain latitude in the initial date of this separation: 330, the foundation of Constantinople; 395, the separation of the Greek part of the empire from its Latin sector; or, finally, the reign of Justinian (527–565). The transition from the earlier to the later art discussed in the next sections took place at different times in different locations; therefore, there can be no precise chronological boundary. Only after Justinian's reign did many Eastern regions submit to the ascendancy of the art of Constantinople, following until the 6th and even the 7th century the paths traced by Christian art in its beginnings. In the West the end of Early Christian art is easier to determine. Closely tied to Roman art, it finished with the collapse of the empire at the end of the 5th century. Then, transformed into a multitude of regional art styles, it assimilated various influences from the East and from the barbaric peoples who superseded their Roman masters.

The vague boundaries of this art in time and space make a definition of its character difficult. Its style evolved from the current Greco-Roman art. The new elements lay not in form but in content: places of worship very different from pagan temples, iconography drawn from the Scriptures. As the hold of the church over public and private life grew, these new elements tended to set traditional subjects completely aside. Early Christian art, while deeply rooted in Greco-Roman art, became a new entity, as distinct from ancient art as from that of the Middle Ages. An obvious difference is the absence of monumental public sculpture. Early Christian sculpture was limited to small pieces and private memorials and only gradually became incorporated into ecclesiastical architecture.

Sarcophagi. The imagery of sarcophagi followed an evolution similar to that of the catacomb paintings. The same biblical and Gospel subjects were introduced into pagan or neutral compositions. In the second or third quarter of the 3rd century, the oldest Christian sarcophagi were hardly distinguishable from the pagan. On one at Sta. Maria Antiqua, Rome, a seated philosopher reading a scroll, a praying figure, and a "Good Shepherd" are "Christianized" by the scenes that accompany them on either side: Jonah resting and the Baptism of Christ. Thus, a sarcophagus from the Via Salaria (Rome, Vatican Museums), which represents the same subjects except for the truly Christian scenes, can be called "Christian" only with reservation.

During the 4th century this iconography was enriched and became more strictly narrative; the miracles of Christ, fully described, were included, the crossing of the Red Sea

Further development of Christian iconography on carved sarcophagi

was often depicted in a long frieze, and the episodes of the Passion of Christ—his arrest, his trial before the Jewish council, his presentation to Pilate, and the Way of the Cross—often extended along the faces of the sarcophagi. The Crucifixion itself was represented by only a bare cross, surmounted by a crown enclosing the monogram of Christ: thus, the symbolic image of the triumph over death. This hesitation to portray the dead Christ on the Cross, an ignominious mode of punishment reserved by the Romans for slaves and abject criminals, disappeared only gradually during the course of the 5th and 6th centuries.

The largest group of Early Christian sarcophagi was found in Rome and its vicinity, although others were found elsewhere in the Mediterranean region. The classicizing style of the first half of the 3rd century became vulgar and a little crude around 300, but it became progressively refined in the time of Constantine and his sons. To the years from 340 to 370 belong the best Roman works: the sarcophagi called the "Two Brothers" (Museo Cristiano), that of Junius Bassus, dated 359, another with columns (both in the grotto of St. Peter's, Rome), that of the "Three Good Shepherds" (Vatican Museums), and, finally, one in S. Sebastiano, Rome, which contains several rare scenes from the story of Lot. While bearing witness to a renaissance of Classical style, they are laden with a new spirituality. A final flourishing occurred near the end of the 4th century in Milan with the decoration of a sarcophagus (S. Ambrogio), which combined an elegant finesse in the figures (due probably to Greek influence) to the vigour of the Roman style.

The sarcophagi of the Middle East and of Ravenna belong principally to the 5th and 6th centuries and to a different artistic tradition. Those of Constantinople and of Asia Minor are fewer in number and lack stylistic homogeneity. Several examples (*e.g.,* sarcophagus of a child and another of the Apostles, end of the 4th century, both in the Arkeoloji Müzeleri, Istanbul) have a harmonious beauty inspired by Classical Greek art; others are in a totally different and more popular style. The sarcophagi of Ravenna, which first appear at the end of the 4th century, stand midway between the Greek art of the East and Latin art. That of Bishop Liberius (4th–5th centuries) of Ravenna at the church of S. Francesco is close to the classicizing Roman sarcophagi in the handling of figures, while the composition—Christ and the Apostles isolated under arcades—finds its models in Asia Minor. Successive waves of Eastern influence affected local style, producing in the 5th century an art distinct from that of the rest of Italy and the Middle East.

Ivory carving. The Christianization of the decorative arts was a slower process than that of monumental art. The presence of pagan imagery on small, movable objects, usually intended for secular use, was less shocking than the same imagery would be on the walls and floors of religious buildings. Because many of these objects were made of precious materials, most of them have disappeared. Only ivories are preserved in considerable number. On a small coffer from Brescia (Civico Museo Romano), second half of the 4th century, Gospel scenes cover the four sides and the top, surrounded by a border of biblical subjects similar to those whose presence has been noted in the paintings of the catacombs and on the sarcophagi. The figures are characterized by a gentle beauty and are close to those of certain Roman sarcophagi of the middle and third quarter of the 4th century. Ivories such as the holy women at the tomb and the Ascension of Christ in the Museo d'Arte Antica, Castello Sforzesco in Milan and in the Bayerisches Nationalmuseum in Munich (Figure 36); six miracles of Christ, divided between the two leaves of a diptych, now in Berlin and in Paris; a coffer in London (British Museum) that bears one of the oldest, if not the oldest, representations of Christ on the cross; and a reliquary found at Pula, Istria, Croatia—all belong to a group of ivories that were produced either in Rome or in northern Italy from the end of the 4th to the middle of the 5th century. In the second half of the 5th century the quality of ivory carving declined in the West; it improved, however, at Constantinople and perhaps other eastern cities, such as Antioch and Alexandria. (He.S.)

Ivories in classical style

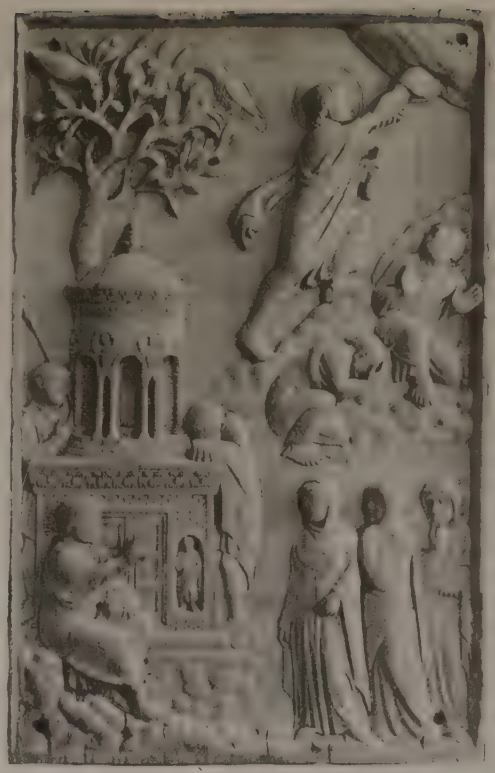


Figure 36: "The Woman at the Tomb and the Ascension," ivory, c. 400. In the Bayerisches Nationalmuseum, Munich. 18.7 cm × 11.7 cm.

By courtesy of the Bayerisches Nationalmuseum, Munich

The Middle Ages

EASTERN CHRISTIAN

The Byzantine era really began with the transference of the capital of the Roman Empire from Rome to the site of ancient Byzantium on the Bosporus in the year AD 330, the new capital thereafter being called Constantinople, after its founder, the emperor Constantine I. Constantine had 17 years earlier been responsible for recognizing Christianity, and from the outset he made it the official religion of the new city. The art dedicated to the service of the faith, which had already begun to develop in the days when Christians were oppressed, received official recognition in the new centre and was also subjected to a number of new influences, so that it owed a debt on the one hand to Italy and Rome and on the other to Syria and Asia Minor, where Oriental elements were prominent. It must not be forgotten that the population of Constantinople and its neighbourhood was Greek, not Latin, so that the poetic and philosophical outlook of the Greek world was itself a very considerable influence.

Constantinople and the Byzantine Empire. Sculpture underwent changes very similar to those in architecture. The decorative work in Hagia Sophia illustrates its nature. In the Classical world naturalistic representation had prevailed; at Hagia Sophia the forms are still basically representational, but they are treated in an abstract manner, more advanced in degree than at St. Polyeuktos. Capitals of the period are similarly stylized even when they use bird or animal forms, for these are usually treated as part of an overall balanced pattern. With this tendency toward stylization in architectural sculpture, it is not surprising to find that three-dimensional, representational sculpture was progressively going out of fashion. Portrait sculptures had been made of most of the early emperors, and the texts report that a mounted figure of Justinian I topped a column in front of Hagia Sophia. But that was the last of the series; figural compositions in high relief had adorned sarcophagi, and similar reliefs had found a place on the walls of churches, but virtually none of these dates from later than Justinian's reign. Instead, flat slabs with low-

Developing stylization in architectural sculpture

relief ornament akin to that on the capitals and cornices of Hagia Sophia, some of it even purely geometric, came into vogue. These slabs were used for the lower sections of windows or to form a screen between the body of the church and the sanctuary; they were later to develop into the high structures called iconostases, which eventually became universal in Orthodox churches.

Ivories. The minor sculptural arts are essential to any treatment of medieval sculpture in general, partly because more is known about them and partly because some of the most able masters of the period preferred to work on small-scale objects, and patronage was ready to support them. Most important are the ivories. They comprise a wide variety of types, ranging from small pyxides—circular vessels used in the liturgy—to large-scale works made up of a number of separate panels, like the famous throne of Maximian (Figure 37), the Archbishop of Ravenna, at Ravenna (*c.* 550; Museo Arcivescovile, Ravenna). Most usual, however, were the flat plaques used as diptychs, book covers, etc. Considerable numbers of these, dating mostly from the late 5th and early 6th centuries, have been preserved. After about the middle of the 6th century, however, ivories become rarer: very few can be dated to the period between the reign of Justinian and the revival of Byzantine art in the 9th century.

Diptychs, or two-panel ivories, seem to have been very popular both for use as book covers and for ceremonial purposes. The most impressive of them were imperial. In these each leaf was made up of five panels; on the central one was a portrait of the emperor; at the sides were standing figures of the consuls; below were scenes, usually of tribute bearers; and above were angels upholding a bust of Christ. They thus illustrated the Byzantine ideas of hierarchy, Christ above and the world below, dominated by the emperor as Christ's vice-regent. The finest of them, known as the Barberini ivory (Figure 38, left), is in the Louvre and probably depicts Anastasius I (491–518); another, of his wife, the empress Ariadne, is divided between several collections.

Figure 37: Ivory throne of Maximian, Archbishop of Ravenna, *c.* 550. In the Museo Arcivescovile, Ravenna. 149.9 × 60.0 cm.
Hirmer Fotoarchiv, Munchen

More numerous today are the diptychs that were issued by the consuls on coming to office. Their fabrication ceased when the office of consul was abolished by Justinian in 541; though by no means are all the consuls portrayed before that time, leaves of the diptychs issued by a large number of them survive. Each leaf consisted of a single plaque. The earlier ones, like that of Probus (408), are still Roman in style; but those dating from just before and just after 500, which constitute the majority, are in a different style, either more ornate or very much simpler. The more elaborate ones are well represented by leaves of the consul Flavius Anastasius (517), in the Cabinet des Médailles, Bibliothèque Nationale, Paris (Figure 38, right); they show the consul enthroned, with lively circus scenes below. The plainer type is represented by a consular diptych of Justinian dated 521 (six years before his accession as emperor), now exhibited in the Castello

Consular diptychs

By courtesy of (left) the Louvre, Paris; photograph (right) Hirmer Fotoarchiv, Munchen

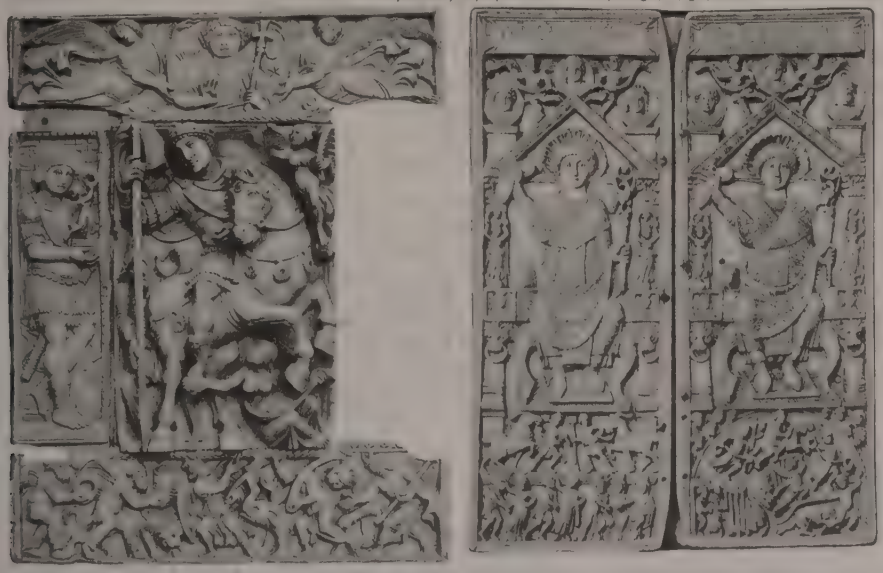


Figure 38: *Early Byzantine ivories.*
(Left) Leaf of the Barberini diptych showing mounted emperor with tribute bearers, *c.* 500. In the Louvre, Paris. 35.0 × 26.7 cm. (Right) Diptych of Flavius Anastasius with the consul enthroned and circus scenes below, 517. In the Cabinet des Médailles, Bibliothèque Nationale, Paris, 35.6 × 25.4 cm.

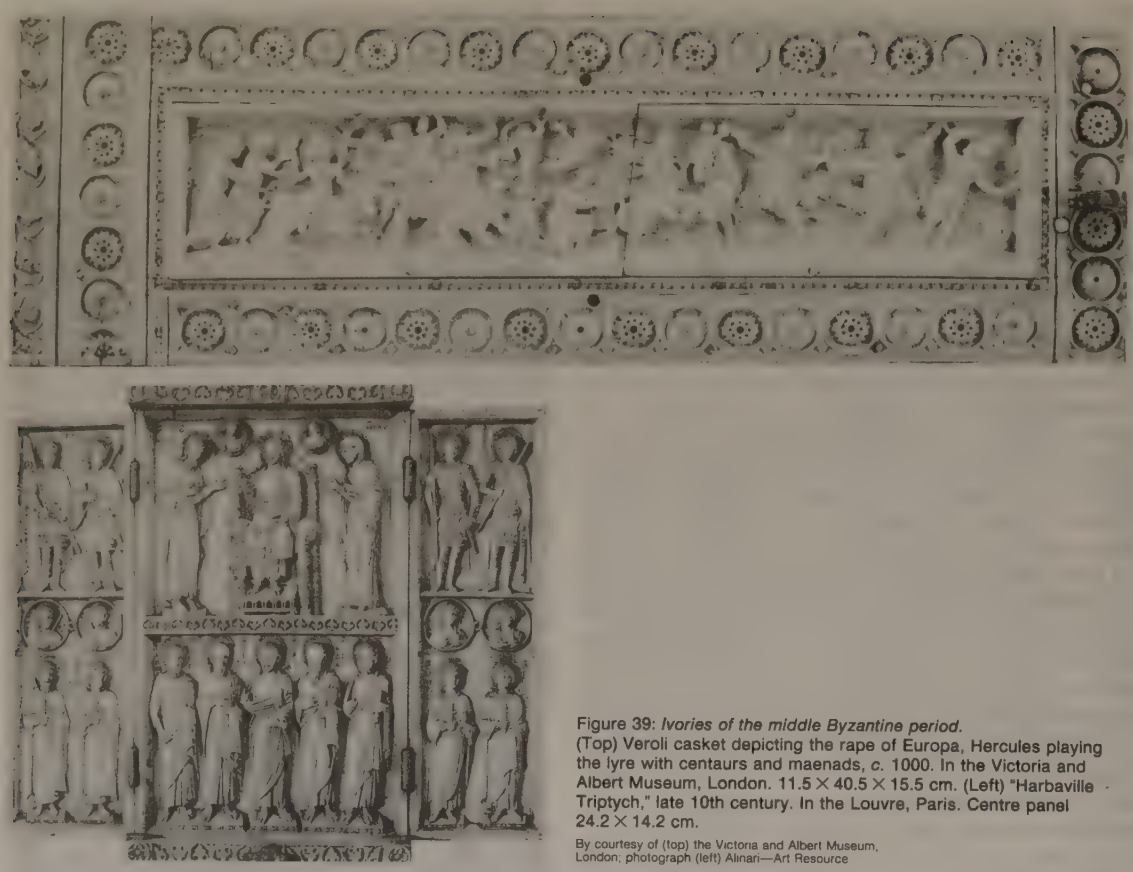


Figure 39: *Ivories of the middle Byzantine period.*
(Top) Veroli casket depicting the rape of Europa, Hercules playing the lyre with centaurs and maenads, *c.* 1000. In the Victoria and Albert Museum, London. 11.5 × 40.5 × 15.5 cm. (Left) "Harbaville Triptych," late 10th century. In the Louvre, Paris. Centre panel 24.2 × 14.2 cm.

By courtesy of (top) the Victoria and Albert Museum, London; photograph (left) Alinari—Art Resource

Sforzesco at Milan, where the decoration is confined to rosettes at the four corners and a medallion with a Latin inscription at the centre.

Most of the official ivories were probably carved at Constantinople, but it seems likely that others, which were intended for more general use or for the church, may well have been done elsewhere. Rome, Milan, Alexandria, and Antioch in Syria were all important centres, and there has been a good deal of dispute among experts as to where many of the ivories were made. Maximian's throne, the most elaborate of them all, has been assigned to Alexandria, Constantinople, and even to Ravenna itself; and there has been argument as to whether the consular diptychs were carved at Constantinople, Rome, or Alexandria. There is, however, unanimity with regard to certain types. Thus, a number of rather small plaques bearing decorations in a clumsy but expressive style can safely be assigned to Palestine, and probably to Jerusalem; another group, characterized by a similar search for realism but by greater technical proficiency, can perhaps be attributed to Antioch. A leaf in the British Museum, with the Adoration of the Magi above and the Nativity below, illustrates the first type; a composite diptych used as a book cover, now at Ravenna, represents the second. Each of its leaves is made up of five panels, like those of the imperial diptychs, but here Christ occupies the central one, and there are scenes from the Gospels and the Old Testament all around.

Alexandrian school

Work of a more polished type, where classical scenes, single figures, or, less often, events from the Bible are the subjects, has been associated with Alexandria. At one time this city was regarded as the primary centre of production, and numerous ivories of major importance were attributed to it, notably the throne of Maximian. The panels that compose the latter are in various styles and are certainly not all of the same school. Those on the sides, depicting scenes from the life of Joseph, are vivid and expressive, whereas those on the front, showing John the Baptist, prophets, and ornamental scrollwork, are grand and elegant. It is possible that the artist who did the Joseph scenes was trained in Alexandria, but most of the rest of the work is now generally regarded as Constantinopolitan, and it was probably there that the throne was carved, wherever the craftsmen had been trained. Also typical of Constantinople, especially during the rule of Justinian, is a large panel in the British Museum representing the archangel Michael. The treatment of this youthful figure and his drapery is in a style reminiscent of classic Greek art, but this is happily combined with ornate decoration and a hieratic composition.

A few ivories bearing secular scenes may also be assigned to the capital; one of the most important is a diptych in the Hermitage at St. Petersburg with depictions of animal combats in the circus.

A fragment of a sceptre in the name of Leo VI (886–912) at the Staatliche Museen Preussischer Kulturbesitz, Berlin, a panel showing the crowning of Constantine VII Porphyrogenitus by Christ (944) in Moscow, and one with the crowning of Romanus II (945) in the Cabinet des Médailles, Paris, can be dated exactly. But in most cases, dates can only be suggested on the basis of style. The ivories have been classified under a number of headings in a monumental survey made by A. Goldschmidt and K. Weitzmann. They term their first group that of Romanus and associate a number of ivories with that showing his crowning, mentioned above; they include triptychs with the deesis on the central panel in the Vatican, the Palazzo Venezia at Rome, and the Louvre, the last known as the "Harbaville Triptych" (Figure 39, left), as well as panels at Dresden, Venice, Vienna, and elsewhere.

Classification of mid-Byzantine ivories

Goldschmidt and Weitzmann's second group is built up around an ivory in the church of Sta. Francesca at Cortona, Italy, which bears the name of Nicephorus II Phocas (963–969). It includes among others a fine triptych with the Virgin on the central panel, at Luton Hoo, Bedfordshire, England. The faces are broader and heavier than those on ivories of the Romanus group. Other groups are distinguished not so much on the basis of date as by form or style, such as groups termed the "painterly" and the "framed," while a more obvious group is composed of caskets. The majority of examples are dated to the later 10th or earlier 11th centuries, but manufacture of objects

in this group apparently continued at least until the early 12th century, the later ones being either more linear in style, like a panel with the Baptist and four Apostles in the Victoria and Albert Museum, London, or the figures being very much elongated, as in a St. John at Liverpool. High relief and deep undercutting were apparently in special favour early in the 11th century.

Though the caskets were no doubt often carved by the same people who carved the plaques, they constitute an independent group not only because of their form but also because they are nearly all adorned with secular motifs that have been drawn from Classical literature. The panels bearing the scenes are framed in bands adorned with rosettes or sometimes human heads in profile; because of this, the caskets are often termed rosette caskets. The most exquisite in execution, if also mannered in style, is one in the Victoria and Albert Museum known as the Veroli casket (Figure 39, top). A few caskets of different type are also known; one at Florence has the rosette borders, but they frame panels bearing Christ, the Virgin, and saints; one at Troyes, France, has no rosette borders, while its side panels show horsemen of Persian type and, at the ends, phoenixes that are distinctly Chinese. During the later part of the 12th century, soapstone plaques became more common than ivories, probably for economic reasons, but they bore low-relief decorations in a very similar style. (D.T.R.)

Georgia. A distinct Georgian sculptural tradition did not emerge until the advent of Christianity, which stimulated a demand for a large number of carved stone reliefs. The earliest of these were based on Early Christian models. In the 8th and 9th centuries the high-relief figures of Early Christian art gave way to figures rendered in wholly linear fashion. In the 10th and 11th centuries the reliefs became gradually more plastic and expressive until they were again freed, to a considerable degree, from the background. At the same time there was an increasing interest in the disposition of figures in a harmonious design. By the 12th century, however, sculptors were beginning to look more to ornamentation than to figural representation. Repetition of themes characterized most of Georgian sculpture in subsequent centuries. Sculpture of all periods was always smaller than life-size.

Linear and plastic treatment of reliefs

Armenia. The stone construction of Armenian churches lent itself to carved decorations, and architectural sculpture was more extensively used in Armenia than in any other country of the Middle East, except Georgia. The reliefs of the 4th-century hypogeum (a subterranean structure hewn out of rock) at Aghts along with those on numerous funerary stelae (upright slabs of inscribed stone) antedating the Arab conquest exemplify the early stages of stone sculpture. Beginning with the 6th century, and perhaps even earlier, floral and geometric motifs as well as figure representations were carved around the windows of the churches, between the arches of the blind arcades, and on the lintels and the lunettes over the doors. Decorative ornaments became increasingly intricate during the later periods.

The outstanding example in Armenian art of the use of architectural sculpture is the Church of the Holy Cross, built in the early 10th century on the island of Aghthamar in Lake Van (Figure 40); this is the earliest medieval example, either in the East or in the West, of a stone building entirely covered with relief sculpture. Around the dome and on the four facades may be seen a variety of animals, vine and other floral scrolls, and large figures of saints and scenes from the Old Testament. A portrait of King Gagik I Artsruni, offering to Christ a model of the church he had erected, appears on the west facade. Such donor portraits, sometimes carved in the round as at Ani, were one of the characteristic features of the decoration of Armenian churches.

Donor portraits

Coptic Egypt. Strictly speaking, the adjective Coptic, when it is applied to art, should be confined to the Christian art of Egypt from the time when the Christian faith may be recognized as the established religion of the country among both the Greek-speaking and Egyptian-speaking elements of the population. In this sense Coptic art is essentially that reflected in the stone reliefs, wood carvings, and wall paintings of the monasteries of Egypt, the earliest foundations of which date from the 4th and 5th centuries AD. It is, however, common practice to include within Coptic art all forms of artistic expression that, like the so-called Coptic textiles, need have no religious intent or purpose. The term has also been further extended to denote stylistic characteristics that can be traced back to the 2nd and 3rd centuries AD and perhaps earlier.

Figure 40: Church of the Holy Cross, Aghthamar, Lake Van, Armenia (now Turkey), early 10th century.
John Donat

A specifically Christian art was slow in developing: when it did emerge, it was not the product of a school of Christian artists inventing new forms of expression. It continued the style current in the country, evolving from the late antique art of Egypt, in which themes derived from Hellenistic and Roman art may or may not have been given new allegorical significance. There is little direct legacy from the art of pharaonic Egypt either in the style of execution or in the choice of decorative themes. The most obvious survival in Christian iconography is the peculiar looped form of cross derived from the ancient Egyptian writing of the word for life (*ankh*). Less convincing is the connection postulated between the concept of *Maria lactans* (representations of the Virgin nursing her child) and bronze and terra-cotta statues of the ancient Egyptian goddess Isis suckling the infant sun god Horus or between representations of saints on horseback and some late figures of the adult Horus in an identical pose.

Origins in ancient Egyptian and Greco-Roman art

The extent to which Egypt may have exerted a major creative influence on Christian art is uncertain in the absence of material remains of the Christian period from Alexandria, the great metropolis of Egypt from the time of the Ptolemies and a city that played an important and, at times, decisive role in the intellectual life of the early church. A series of Christian ivory carvings, of unrecorded provenance, is frequently referred to as Alexandrian on stylistic considerations and adduced as proof of a continuing artistic skill in the Hellenistic tradition.

Objects found in the hinterland depart from the Classical canons of proportion and mode of representation. Political and economic conditions in Egypt from the time of its incorporation in the Roman and, later, Byzantine empires

doubtless account for much of the provincial appearance of Egyptian and Coptic art and the emergence of a freer, more popular folk style. Lack of the kind and degree of patronage that had been given by the pharaohs, Ptolemies, and, to some extent, Roman emperors to the old religion of Egypt meant an impoverishment of schools of skilled craftsmen, avoidance of costlier materials, and a decline in the high standard of finish. Particularly noticeable is the absence of carving in the round, of work of monumental scale, and of the use of the harder ornamental stones that had been characteristic of pharaonic art.

Characteristic Coptic stylistic features are to be observed in tombstones from the Delta site of Terenuthis. These depict the dead man frontally posed beneath a gabled pediment of mixed architectural style, hands extended at right angles from the body and bent upward from the elbow in the orans (praying) position, a pose that appeared frequently in the earliest Christian art in Rome. There is no firm evidence, however, that the community was Christian. Similarly, the series of architectural elements carved in relief from Oxyrhynchus and Heracleopolis may not all be from Christian buildings. The earlier material from Heracleopolis, dating probably from the 4th century, is notable for its figure subjects drawn from classical mythology, carved in a deep relief that leaves them almost freestanding, producing an effective play of light and shade. As such reliefs were painted, the absence of fine detail in the carving was less noticeable.

Much of the material available for a study of Coptic sculpture has not been found in context, and, in the absence of assured information concerning its provenance and of circumstantial evidence for dating (even in the cases of pieces from known sites), it is impossible to provide a detailed account of the development of Coptic sculpture. In general, the figures are stiff in pose and movement; there is a tendency for the carving to become flat, and there is little in the way of narrative scenes drawn from biblical stories. The most successful carvings are probably the impressive variety of decorated capitals, particularly from the monasteries of Apa Jeremias at Ṣaqqārah and of Apa Apollo at Bāwīṭ. Among them are basket-shaped examples decorated with plaitwork, vine and acanthus leaves, and animal heads. The form imitates a style introduced into Constantinople by the emperor Justinian I, and it is clear that, in the hinterland of Egypt, there was during the 6th century certain artistic influence on Coptic art from Byzantium, despite religious and political differences. Contemporary Byzantine influence seems to have been at work on other architectural elements at Bāwīṭ, as, for example, in the finely carved limestone pilaster depicting, on one side, a geometric and floral pattern surmounted by a saint and, on the other, vine scrolls and birds below an archangel. (A.F.Sh.)

Carved capitals

WESTERN CHRISTIAN

With the dissolution of the Roman Empire in the West, cultural hegemony passed to the Eastern Empire, but older traditions remained in western Europe and intermingled with several invaders—Germanic tribes arriving from the north and Christians arriving from Constantinople as well as from Rome. The Merovingian art of the Franks, which was culturally predominant throughout Europe in the 6th century, survives principally in grave relics, such as jewelry, hollowware, and the like.

In Italy the Lombards, who invaded the country in 568, propagated Germanic art, but there is a strong Mediterranean influence in the sculpture—stone plaques for choir screens, altars and altar canopies, sarcophagi, and details of architecture, for example; the abstract decorations, many of them interlaced motifs, were to be blended with more and more Byzantine elements (Figure 41). The creatures and vegetation become almost impossible to recognize—they aspire, as it were, to be ornamental stone writing rather than representation. Similar ornaments were also applied in stucco; for example, in S. Salvatore at Brescia and especially in the famous Tempietto at Cividale del Friuli (both 8th century). At Cividale del Friuli, standing figures of saints have been incorporated in decoration in which the Byzantine influence is obvious.

Figure 41: Marble relief inscribed by the Patriarch Sigwald, part of the canopy over the baptismal font in the cathedral of Cividale, Italy, 762–776. 0.91 m × 1.52 m.
Bildarchiv foto Marburg—Art Resource

In Ireland, monumental crosses represented the Celtic Christian tradition, and similar Anglo-Saxon crosses may be found in England (Figure 42). The abstracted decoration recalls the relief style in Italy, but here the surface is not a flat plane but is packed with round, knoblike projections that create a plastic rather than a glyphic effect.

Carolingian and Ottonian periods. The cultural revival of the Carolingian period (768 to the late 9th century), stimulated by the *academia palatina* at Charlemagne's court, is the first phase of the pre-Romanesque culture, a phase in which late Classical and Byzantine elements amalgamated with ornamental designs brought from the East by the Germanic tribes. The German Ottonian and early Salian emperors (950–1050), who succeeded the Carolingians as rulers of the Holy Roman Empire, assumed initially the Carolingian artistic heritage, although Ottonian art later evolved into a distinct style.

By courtesy of the Irish Tourist Board

Figure 42: "Cross of Muiredach," Monasterboice, Ireland, 923. Height 5.38 m.

Little Carolingian sculpture has survived, but in Ottonian days the sculpting of freestanding statues was taken up again, although the earliest specimens, serving as they did as reliquaries, were still closely related to the silversmith's and goldsmith's art; for example, the famous statue of "Sainte-Foy" at Conques (France) and the "Golden Madonna" at Essen. The wooden "Gero Crucifix" (about 73.6 inches [187 centimetres] high; cathedral of Cologne; Figure 43), which was carved before 986, already reveals a certain realism in the representation of the shape of the body, in contrast to the contemporary crucifix of Gerresheim (before 1000). The so-called Bernward Crucifix at Ringelheim (Germany) is between the two. The reliefs on the wooden doors of Sankt Maria im Kapitol at Cologne display an affinity with the mid-11th-century Romanesque ivories of the Meuse district. The Carolingian bronze doors in Aachen were imitated at Mainz, where Bishop Willigis had similar portal wings made for his cathedral. He was far surpassed, however, by Bernward at Hildesheim, who had the still extant door wings of the cathedral (1015) decorated with typological images in parallel, scenes from the Old and the New Testament; in theme, the images go back to early Christian examples Bernward had seen in Italy, but the force of the gestures and the use of unadorned surface as dramatic interval in the episode of Adam and Eve reproached by the Lord has no precedent in the history of art (Figure 44). The influence of Classical art manifests itself clearly in the so-called Christ's Column (12.8 feet [3.9 metres] high; *c.* 1020; St. Michael's, Hildesheim), which, with its figures spiralling around the shaft, reminds one of the triumphal columns of Trajan and Marcus Aurelius. Originally, it was crowned by a cross. As belonging to the art associated with Bernward, one must also reckon the seven-branched candlestick in the Minster of Essen (90.6 inches [230 centimetres] high; before 1011) and the bronze crucifix at Essen-Werden (42.5 inches [108 centimetres] high; *c.* 1060), a late product of the same school.

"Christ's Column"

Figure 43: The "Gero Crucifix," carved oak corpus (with contemporary nimbus and stem), before 986. In the cathedral of Cologne. Height 187 cm.

Bildarchiv foto Marburg—Art Resource/EB Inc

Romanesque. The term Romanesque—coined in 1818—denotes in art the medieval synthesis of the widespread Roman architectural and artistic heritage and various regional influences, such as Teutonic, Scandinavian, Byzantine, and Muslim. Although derived primarily from the remains of a highly centralized imperial culture, the Romanesque flowered during a period of fragmented and unstable governments. It was the medieval monasteries, virtual islands of civilization scattered about the continent, that provided the impetus—and the patronage—for a major cultural revival.

The bronze "Christ's Column" is a modest prophecy of the monumental spirit that would distinguish the sculptural decoration of the new monastic buildings rising in much of western Europe. Developed in the abbey doorways and on the pillars and capitals of cloisters, where the sculptor had to learn anew the technique of stone carving and of rendering the human figure, this spirit gradually grew stronger.

During the 11th century more and more churches were constructed in the Romanesque style, the massive forms of which are another indication of this sculptural instinct. Romanesque sculpture culminated in France in the great semicircular relief compositions over church portals, called tympanums. The example at Moissac (*c.* 1120–30), which represents the Apocalyptic vision with the 24 elders, is a particularly brilliant demonstration of how devices of style can so transform the objects of nature that they seem entirely purged of terrestriality (Figure 45). All the forms are suspended in a predominating plane that denies physical space. Differences in scale are masterfully exploited: the tiny figures of the elders are a foil to the looming image of Christ in the centre. With great consistency, every detail has been subjected to a process of stylization that produces rhythmic patterns in the drapery, hair, and feathers. The central figure is so flattened as to appear disembodied, while the two towering angels have been so attenuated that their bodies have lost all mass.

The astonishing variety that master sculptors such as Gislebertus, Benedetto Antelami, and Nicola Pisano achieved within the confining principles of Romanesque

Bildarchiv foto Marburg—Art Resource/EB Inc

Figure 44: "Adam and Eve Reproached by the Lord," bronze panel from the doors of Bishop Bernward at the cathedral, Hildesheim, West Germany, 1015. Panel 58.6 cm × 196 cm.

Figure 45: "Christ of the Apocalypse, with the 24 Elders," tympanum of the south portal of the abbey church of Saint-Pierre at Moissac, France, *c.* 1120–30.
Yan

style can be illustrated, on the one hand, by the tympanums of Burgundy, such as the spectral "Last Judgment" at Autun (Figure 46) or the "Pentecost" at Vézelay, and, on the other, by the less visionary sculpture of Provence, such as that of Saint-Trophime in Arles (Figure 47) or of the church in Saint-Gilles, which retain many of the forms and characteristics of Classical antiquity.

Another sculptural form that reappeared in Europe during the latter part of the Romanesque period was sepulchral sculpture, in which a sculptured figure of the deceased was cut or molded on top of a sarcophagus or on the sepulchral slab set into the floor of an abbey or cloister. (J.J.M.T./Ed.)

Gothic

Origin of the term Gothic

The difficulty with many anatomies of Gothic art is that they become involved in attributing a meaning to Gothic that it is incapable of sustaining. It is not, for one thing, a medieval word; instead, it is an invention of the 16th century attributed, as it were, posthumously, by historians after the Gothic style had been trampled into virtual insensibility by the Italian Renaissance. The word refers to the Teutonic tribes who were thought to have destroyed Classical Roman art and were thus considered barbarians. But nobody in the 13th century thought of himself as Gothic. The fact is that the literature of art criticism is virtually nonexistent in the Middle Ages. Certainly people talked about art, patrons valued it, connoisseurs appraised it. But the terms in which this was done must now, for the most part, be a matter of speculation or imagination. There was not necessarily anything mysterious about this. It is common to suppose that medieval discussions on art were infused with a degree of spirituality. This is probably mistaken. There is, for instance, little that is spiritual about financing the building of a gigantic cathedral. It is certain that clergymen preached sermons about art, giving it a spiritual and symbolic interpretation. It is also true that, since a large proportion of art served a religious function, artists were, in some sense, "servants of God." But they were also the servants of far more worldly considerations, such as earning a living or achieving a reputation, and these should never be discounted in any imaginative re-creation of the medieval artist's existence.

Bildarchiv foto Marburg—Art Resource/EB Inc

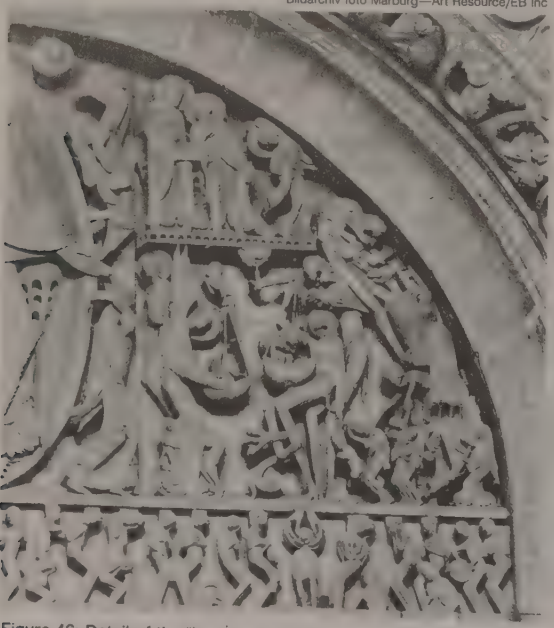


Figure 46: Detail of the "Last Judgment," from the west tympanum of the cathedral of Saint-Lazare, Autun, France, carved by Gislebertus before 1135.

Giraudon—Art Resource

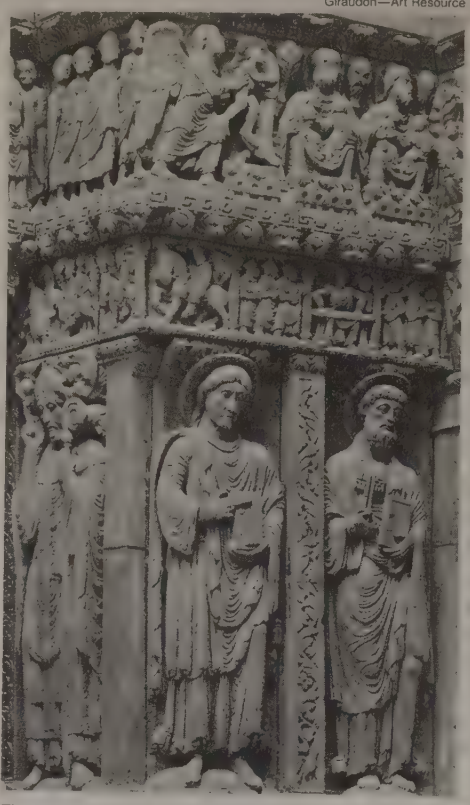


Figure 47: Detail from the main portal of the church of Saint-Trophime, Arles, France, 12th century.

EARLY GOTHIC

Portal sculpture

Throughout this period, as in the Romanesque period, the best sculptors were extensively employed on architectural decoration. The most important agglomerations of figure work to survive are on portals, and, in this, once again, the church of Saint-Denis assumes great significance. The western portals (built 1137–40), part of a total facade design, combined features that remained common throughout the Gothic period: a carved tympanum (the space within an arch and above a lintel or a subordinate arch); carved surrounding figures set in the voussoirs, or wedge-shaped pieces, of the arch; and more carved figures attached to the sides of the portal. As it survives, Saint-Denis is disappointing; the side figures have been destroyed and the remainder heavily restored. The general effect is now more easily appreciated on the west front of Chartres cathedral.

If one compares the portals here (*c.* 1140–50; Figure 48, left) with those of early 13th-century Reims, one can see that the general direction of the changes in this early

Figure 48: *French early Gothic architectural sculpture.*
(Left) Figures from the Old Testament, centre portal of the west front of Chartres cathedral, *c.* 1140–50. (Centre) Saints, south transept, Chartres cathedral, *c.* 1210–20. (Right) Apostles of the Judgment Portal, north transept, Reims cathedral, *c.* 1225.
(Left) Madame Simone Roubier, Paris, (centre) Giraudon—Art Resource/EB Inc. (right) Archives Photographiques

period of Gothic sculpture was toward increased realism. The movement toward realism is not manifest in a continuous evolution, however, but in a series of stylistic fashions, each starting from different artistic premises and achieving sometimes a greater degree of realism but sometimes merely a different sort of realism. The first of these fashions can be seen in the sculpture on the west front of Chartres. That the Christ and the Apostle figures are in some sense more human than the Romanesque apparitions at Vézelay and Autun (*c.* 1130) need hardly be argued. That the figures, with their stylized gestures and minutely pleated garments, are at all "real" is doubtful. That their forms are closely locked to the architectural composition is clear. The features of the Chartres sculpture had a wide distribution; they are found, for example, at Angers, Le Mans, Bourges, and Senlis cathedrals. There are stylistic connections with Burgundy and also with Provence. The fashion lasted from *c.* 1140 to 1180.

The centre of development for the second style lay in the region of the Meuse. The activity of one of the chief artists, a goldsmith called Nicholas of Verdun, extends at least from the so-called Klosterneuburg altar (1181) into the early years of the 13th century. His style is characterized by graceful, curving figures and soft, looping drapery worked in a series of ridges and troughs. From these troughs is derived the commonly used German term for this style—*Muldenstil.* This drapery convention is essentially a Greek invention of the 4th century BC. It seems likely that Nicholas seized the whole figure style as a tool to be used in the general exploration of new forms of realism. It remained extremely popular well into the 13th century. A rather restrained version of the style decorated the main portals of the transepts (the transversal part of a cruciform church set between the nave and the apse or choir) of Chartres (*c.* 1200–10; Figure 48, centre). It is also found in the earliest sculpture (*c.* 1212–25) of Reims cathedral (Figure 48, right) and in the drawings of the *Sketchbook* of Villard de Honnecourt (*c.* 1220).

In the opening years of the 13th century yet another type of realism emerged. It seems to have originated at Notre-Dame, Paris (*c.* 1200), and to have been based on Byzantine prototypes, probably of the 10th century. The looping drapery and curving figures were abandoned; instead, the figures have a square, upright appearance and are extremely restrained in their gestures. Figures in this style are found at Reims, but the major monument is the west front (*c.* 1220–30) of Amiens cathedral (Figure 49, left).

Once again, the style changed. On the west front of Reims worked a man called after his most famous figure, the Joseph Master (Figure 49, right). Working in a style that probably originated in Paris *c.* 1230, he ignored the restraint of Amiens and the drapery convolutions of the *Muldenstil* and produced (*c.* 1240) figures possessing many of the characteristics retained by sculpture for the next 150 years: dainty poses and faces and rather thick drapery hanging in long V-shaped folds that envelop and mask the figure.

The Joseph Master

Another aspect of this quest for realism was the spasmodic fashion throughout the 13th century for realistic architectural foliage decoration. This resulted in some astonishingly good botanical studies—at Reims cathedral, for example.

The effects elsewhere in Europe of this intense period of French experiment were as piecemeal and disjointed as the effects of the architectural changes. In England, the concept of the Great Portal, with its carved tympanum, voussoirs, and side figures, was virtually ignored. The remains of a portal the style of which may be connected with Sens cathedral survive from St. Mary's Abbey, York, England (*c.* 1210). Rochester cathedral (*c.* 1150) has carved side figures, and Lincoln cathedral (*c.* 1140) once had them. The major displays of English early Gothic sculpture, however, took quite a different form. The chief surviving monument is the west front of Wells cathedral (*c.* 1225–40), where the sculpture, while comparing reasonably well in style with near-contemporary French developments, is spread across the upper facade and hardly related at all to the portal.

In Germany, the story is similar. On the border between France and Germany stands Strasbourg, the cathedral of which contains on its south front some of the finest sculpture of the period (*c.* 1230). A very fine and delicate version of the *Muldenstil,* it comes reasonably close to the best transept sculpture of Chartres. But it differs in two important respects. Predictably, its architectural framework is entirely different; and it has the slightly shrill emotional character, common in German art, that represents an effort to involve and move the spectator. Shrill emotionalism is again found at Magdeburg cathedral in a series of "Wise and Foolish Virgins" (*c.* 1245) left over from some abandoned sculptural scheme. Influenced by Reims rather than Chartres, the sculpture of Bamberg cathedral (*c.* 1230–35; Figure 50, left) is a heavier version of the *Muldenstil* than that at Strasbourg. But of all this German work, by far the most interesting complex is in the west choir (*c.* 1250) of Naumburg cathedral (Figure 50, right). Here, the desire for dramatic tension is exploited to good effect, since the figures—a series of lay founders in

Figure 49: *Styles of realism in portal sculpture in France.*
(Left) Statue of Christ ("Le Beau Dieu"), centre portal of the west facade, Amiens cathedral, *c.* 1220–30. (Right) Visitation, detail of the Virgin's Portal, west facade, Reims cathedral, 1225–45.
Jean Roubier

contemporary costume—are given a realistic place in the architecture, alongside a triforium gallery. Naumburg also has a notable amount of extremely realistic foliage carving.

It is hard to say what a French mason would have made of this English and German work. With the major Spanish work of the period, however, he would have felt instantly at home. Burgos cathedral has a portal (1230s) that is very close to the general style of Amiens, and its layout is also, by French standards, reasonably conventional.

HIGH GOTHIC

Late sculptural developments of the early Gothic period were of great importance for the High Gothic period. The Joseph Master at Reims and the Master of the Vierge Dorée at Amiens both adopted a drapery style that, in various forms, became extremely common for the next century or more; both introduced into their figures a sort of mannered daintiness that became popular. These features appear in an exaggerated form in some of the sculpture for the Sainte-Chapelle, Paris.

On the whole, this period saw the decline of architectural sculpture. Given the emphasis placed on geometric patterning by the Rayonnant style, perhaps this is not surprising. A few portals, such as those on the west front of Bourges cathedral, were completed, but they have a very limited interest. The field of sculpture that expanded with great rapidity was the more private one, represented by tombs and other monuments.

For this, the family feeling of Louis IX was partly responsible. By making sure that both his remote ancestors and his next of kin got a decent burial—or reburial—he was responsible for an impressive series of monuments (the remnants of which are now chiefly in Saint-Denis) executed mainly in the years following 1260. Although earlier examples and precedents may be found, Louis IX had a large share in popularizing the idea of the dynastic mausoleum, and many other important people followed suit.

The monuments executed for St. Louis have come down in such a battered state (almost entirely as a result of the destruction wrought during the French Revolution) that it is difficult to generalize about them. One can say, however, that Louis's masons popularized two important ideas. One was the tomb chest decorated with small figures in niches—figures generally known as weepers, since they often represented members of the family who might be presumed to be in mourning. Later, in the early 14th century, the first representations appear of the heavily cloaked and cowled professional mourners who were normally employed to follow the coffin in a funeral procession. The second innovation introduced by Louis's masons lay in the emphasis given to the effigy. Around 1260 the first attempts were made to endow the effigy with a particular character. This may not have involved portraiture (it is obviously hard to be sure), but it did involve a study of different types of physiognomy, just as the botanical carving of the early Gothic period had involved a study of different kinds of leaves.

A somewhat similar story may be told of English sculpture during this period. The architectural carving found at Westminster Abbey (mainly of the 1250s) has much of the daintiness of contemporary French work, although the drapery is still more like that of the early Chartres or Wells sculpture than that of the Joseph Master. The baggy fold forms of the Joseph Master rarely appear in England before the sculptured angels of the Lincoln Angel Choir (after 1256).

Architectural sculpture in England probably remained more interesting than the continental equivalent because first-rate masons continued to work in this field in England until the end of the 13th century. Hence, around 1295 one can still find a work such as the botanical carving of Southwell Chapter House. Even in the 14th century, there are such architectural and sculptural curiosities as the west front of Exeter cathedral. Sculptural interest, however, in buildings such as Gloucester Cathedral Choir (begun soon after 1330), where the effect depends on traceried panels, is virtually nonexistent; and the "leaves of Southwell" were succeeded almost at once by an extremely dull form of foliage commonly known as "bubbleleaf,"

Figure 50: *German Gothic sculpture.*
(Left) "Bamberg Horseman," possibly a king or emperor, Bamberg cathedral, Germany, *c.* 1230–35. (Right) "Ekkehard and Uta," Naumburg cathedral, Germany, *c.* 1250.
(Left) Foto Marburg—Art Resource/EB Inc., (right) H. Roger-Viollet

which remained more or less standard for the 14th and 15th centuries.

Monuments of Westminster Abbey

As in France, much of the virtuosity in carving went into private tombs and monuments. The best surviving medieval mausoleum is Westminster Abbey, where a large number of monuments in a variety of mediums (especially purbeck, bronze, alabaster, and freestone) is further enhanced by some of the floors and tombs executed by Italian mosaic workers introduced by Henry III. Especially well preserved is the tomb of Edmund Crouchback, earl of Lancaster (died 1296), which has a splendid canopy and retains some of its original colouring.

As in the early Gothic period, the west of England produced some highly original work that appears to stand outside the normal canon of European development. The earliest monument in this series is the tomb of Edward II (*c.* 1330–35), which is notable for one of the most elaborate surviving medieval canopies. It is preceded stylistically by the wooden canopies of stalls in Exeter cathedral and thus is likely to be a translation into stone of carpenters' work. It was followed by a series of monuments, in Tewkesbury and elsewhere, extending into the 15th century and then dying out.

German High Gothic sculpture is represented by some rather dainty, elegant figures, enveloped in curving and bulky drapery, around the choir of Cologne cathedral (consecrated in 1322). There is also some impressive figure sculpture on the west front of Strasbourg cathedral (begun after 1277). It is strongly influenced by the Joseph Master of Reims but also by the earlier Gothic sculpture of Strasbourg itself. Although it varies in style, much of it is far more expressive than the related French work. The sculptors seem to have been trying to capture an emotive mood.

Spanish High Gothic architectural sculpture is probably less interesting but, by French standards, is more conventional than the German. Major portals exist at León (13th century) and Toledo (14th century) cathedrals, which conform more or less to the rather elegant and mannered French style. Spain also possesses a considerable number of interesting tombs from this period.

Italian Gothic. The figurative arts in Italy during the period 1250–1350 have a strong line of development. The most important 13th-century sculptors were Nicola Pisano (1210/20–1278/84) and his son Giovanni (*c.* 1245–after 1314). Both worked mainly in Tuscany, and both executed pulpits that rank as their major completed works. Nicola's style, as seen in the Pisa Baptistery (1259–60) and Siena cathedral (1265–68) pulpits, was heavily influenced by Classical sculpture—especially by the facial types and the methods of constructing pictorial relief compositions (Figure 51). Nevertheless, his reliefs resemble 13th-century sculpture, particularly in the handling of the drapery. Moreover, in moving from Pisa to Siena, one is conscious of a transition from a strongly antique style to something much closer to northern Gothic sculpture. Nicola's use of Classical ideas was in some way linked with a search for a more realistic style. It forms, in this respect, an interesting parallel to the *Muldenstil* work of Nicholas of Verdun, who was active in the Mosan region from the late 12th to the early 13th century.

The sculptural style of Giovanni does not develop from that of his father. His pulpit in S. Andrea Pistoia (completed 1301), for instance, is technically less detailed and refined but emotionally much more dramatic. While it is possible that the emotionalism of his work was inspired by Hellenistic sculpture, it is also possible that Giovanni had travelled in and been influenced by the north, especially Germany.

Giovanni's first major independent work was a facade for Siena cathedral (*c.* 1285–95). The lower half alone was completed, and it survives in the present building along with a large proportion of Giovanni's imposing figure sculpture. It is quite dissimilar to French facades,

Figure 51: Marble pulpit by Nicola Pisano, 1259–60. In the Pisa Baptistery.
Alinari—Art Resource/EB Inc

although the placing of the main sculpture above the portals finds an elusive parallel in Wells cathedral, in England (*c.* 1225–40).

Influence of the Pisano workshop

The fame of Nicola's workshop spread to other areas of Italy. For S. Domenico in Bologna, his workshop made a shrine for the body of St. Dominic (1260s). And in Milan, a shrine for the body of St. Peter Martyr was made for S. Eustorgio (1335–39) by Giovanni di Balduccio in a style derived from the Pisano workshop. The most famous Pisano "exports," however, were Arnolfo di Cambio, who worked for the papal court in Rome *c.* 1275–1300, and Tino di Camaino, who worked at the Neapolitan court *c.* 1323–37.

Arnolfo's style is the more difficult to understand. Although he worked alongside Giovanni Pisano during the 1260s, their works have little in common. Arnolfo's sculpture is very solid and impassive. He excelled at formal, static compositions, such as were required for church furniture and for tombs. He designed the funerary chapel as well as the tomb of Pope Boniface VIII and like the Pisanos was architect as well as sculptor; indeed, he was the first architect of the new cathedral of Florence (founded 1296).

Tino di Camaino went south after a training in Siena and a successful career in Tuscany. Sometimes his style approaches the elegance and sweetness of northern 14th-century sculpture, but there is generally a residual heaviness, especially in the faces, that reminds one of his origins in the Pisano circle. He was famous as a tomb sculptor, and the largest collection of his monuments is in Naples (much of the sculpture, however, was executed by his workshop). The tombs make an interesting comparison with those of the French and English royal houses. At another mausoleum (of the Scaliger family), at Verona, the figure sculpture is reminiscent of the Pisano style, but the decorative canopy work is more elaborate and closer to northern art.

The workshop of the facade of Orvieto cathedral and the work of the sculptor and architect Andrea Pisano (no relation to Nicola and Giovanni) are less clearly connected with the Pisano tradition. The facade of Orvieto was designed by the Sienese Lorenzo Maitani *c.* 1310. The sculptural decoration is in varying styles, the best of which is an extraordinarily low and delicate relief that gives an almost pictorial quality.

Andrea Pisano is known chiefly through the bronze doors completed for the Baptistery of Florence cathedral during the 1330s. The scenes of the life of St. John the Baptist are set in quatrefoils (a four-lobed foliation), a common High Gothic decorative motif. Within this awkward shape, the episodes are composed with masterly skill. Although nothing certain has been established about the training of Andrea Pisano, his background is likely to have been similar to that of some of the Orvieto sculptors. The main difference is the evident impact of Giotto's painting, which led Andrea to make his figures rather stocky and solid.

Andrea had a son, Nino Pisano, about whom little is known but from whose hand a group of Madonnas survives. They are interesting in that they veer strongly in the direction of daintiness and sweetness and, to this extent, look more northern than almost any other group of Italian sculpture before the early work of Lorenzo Ghiberti.

International Gothic. The plastic arts are harder to understand in this period, because they have been far more frequently the subject of wanton destruction. Enormous quantities, for example, of goldsmiths' work owned by the French royal family have almost entirely vanished. A few of the remaining pieces testify to the quality of the work, which is beautifully finished and gaily coloured in the technique of *en ronde bosse* enamelling—for example, the "Thorn Reliquary" (*c.* 1400–10; British Museum, London), and the "Goldenes Rössel" at the Stiftskirche, Altötting, Germany (1403).

More seriously, large quantities of private monumental sculpture have been lost in France and the Low Countries. The main sculptor of the French royal family in the second half of the 14th century was a native of Valenciennes, André Beauneveu. His reputation was so widespread that he rather surprisingly earned a mention in the chronicles of Jean Froissart. He produced a large number of monuments, especially for King Charles V, of which several effigies survive (Figure 52). This sculpture, while technically good, is somewhat pedestrian and hardly serves as a prelude to the work of Claus Sluter, who worked for Charles V's brother Philip the Bold, duke of Burgundy.

Claus Sluter

Sluter's surviving work is mainly at Dijon, France, where he was active from about 1390 to about 1406. His figure style is very strongly characterized and detailed and, at times, emotional. This suggests that his origins are German and that he may have come from the region of Westphalia. The intrusive realism of Sluter's work, however, is

Giraudon—Art Resource/EB Inc

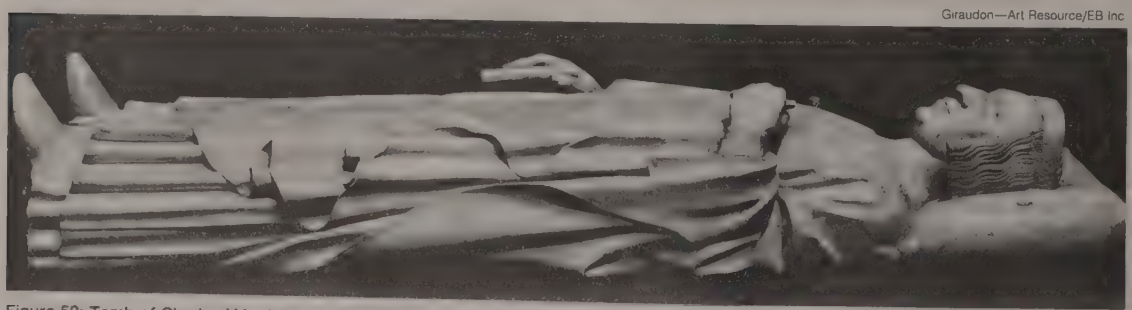


Figure 52: Tomb of Charles V in the abbey of Saint-Denis, France, by André Beauneveu, 14th century.

also symptomatic of a gradual change in sculptural style during this period. The strong characterization of the faces of his figures finds parallels in the near-contemporary triforium busts and Přemyslid tombs in St. Vitus' Cathedral in Prague. Sluter's drapery style, which veers dramatically away from the somewhat reticent elegance of previous court sculpture, also has parallels in the east. Bohemia and Austria possess a series of Madonna figures (*Schöne Madonnen*) swathed in extremely elaborate and artificial drapery arrangements.

The International Gothic sculptural style forms an interesting prelude to developments in Italy, especially to the early work of Donatello and the gradual introduction of Classical ideas into sculpture, for these ideas can be seen as part of a search for an alternative to the elegance of International Gothic. How far Florentines had any knowledge of northern developments is not clear. Ghiberti certainly knew a little about them; moreover, the task of rebuilding Milan cathedral during this period (*c.* 1400) brought large numbers of northern masons across the Alps. As yet, however, the extent to which the sculpture on Milan cathedral was influenced by northern ideas has not been determined.

England stands apart from much of the development represented by Sluter's style. The royal tombs in Westminster Abbey, which extend up to Richard II (died 1400), do not reflect changes subsequent to the phase of André Beauneveu. Further, a fashion for bronze effigies, going back to the effigy of Henry III (1291–93), persisted in England. But whatever the regional idiosyncracies, Westminster tombs, existing as a group in situ, provide a somewhat faded and battered impression of what these great collections of medieval family monuments looked like.

Figure 53: Weepers from the tomb of Philip the Bold, by Claus Sluter and Claus de Werve, completed 1411. In the Musée des Beaux-Arts, Dijon, France. Height 40.6 cm.
R Remy

LATE GOTHIC

In the years around 1400, when International Gothic flourished, Italian and northern artists had achieved some sort of rapprochement. Under the renewed influence of antique art, Italy drew away again, and it was not until the 16th century that the north showed any real disposition to follow suit in the imitation of Classical models. While painting and architecture of the 15th century have a reasonably well-defined development, sculptural development is harder to trace—partly because much crucial work (especially in the Low Countries) has been destroyed. It is clear, however, that elaboration rather than restraint was the rule—indeed, the exceptions to the rule (mainly found in France) stand out. This taste for the highly complicated and elaborate—especially in Spain and Germany—was encouraged by the dual influences of painting and architecture. Like the painters, the sculptors enjoyed giving extremely realistic detail and expression to their figures; and, like the architects, they enjoyed complicated tracery work, often encasing their compositions in tabernacle-like enclosures of brilliantly fantastic architecture. To 20th-century eyes, the result may seem overloaded and the total impression exhausting; but in its time the work of, for example, Michael Pacher or Veit Stoss must have been admired precisely for the way in which the sculptor used every conceivable opportunity to display his virtuosity.

The taste for the elaborate

One interesting characteristic of the late Gothic period deserves comment: the increase in the amount of art produced by foreign artists for countries such as Hungary, Poland, and Scotland. Competition between countries for the work of the best artists was not new. Throughout the Middle Ages artists travelled widely. In the 13th century Villard de Honnecourt went from northern France to Hungary, and Roman marble workers journeyed to Westminster. In the period *c.* 1400 there was much interchange between northern and southern Europe. In the 15th century, this general pattern was confirmed; the Netherlandish sculptor Gerhaert Nikolaus von Leyden, for instance, became court sculptor in Vienna, and the Italian sculptor and architect Andrea Sansovino served the Portuguese court in the 1490s. There is also the work of the Franconian sculptor Veit Stoss for the Polish court at Cracow (*c.* 1480) and the work of Bernt Notke of Lübeck for Aarhus (Denmark), Tallinn (Estonia), and Stockholm (*c.* 1470–90). Numerous other objects could be added. More specifically, there is the altar executed by Meister Francke of Hamburg for Helsingfors (1420s) and Hugo van der Goes' panels for the Palace of Holyrood, near Edinburgh (1470s).

Sluter's work for the court of Burgundy lasted about 15 years. During this time, he worked on three major items: the main portal of the chapel of the Charterhouse near Dijon; inside the chapel, the tomb of his patron, Philip the Bold (Figure 53); and a large Calvary group for the Charterhouse cloisters. When he died in 1406, the continuance of his work was assured by the employment of his nephew and heir, Claus de Werve, until his death in 1439. Further, the pattern of the finally completed tomb of Philip the Bold became famous immediately and was frequently imitated all over Europe.

The forcefulness and boldness of Sluter's sculpted figures is combined with elaborate decorative work—on the canopy of the tomb of Philip the Bold, for example. A similar decorativeness is found in the contemporary carved Dijon altarpieces of Jacques de Baerze. The combination remained more or less constant for the rest of the Gothic period.

The spread of this style is hard to trace. In Germany, the most interesting artists worked in the second half of the century. Two of the more important sculptors were Gerhaert Nikolaus von Leyden and Michael Pacher of Brunico. They were followed by a number of virtuoso southern German artists: Veit Stoss of Nürnberg, Tilman Riemenschneider of Würzburg (Figure 54), and Adam Kraft of Nürnberg. In northern Germany, the most original figure was Bernt Notke of Lübeck. Much of the fantastic decorative involvement of his work may now seem overwhelming. The love of realistic detail is well illustrated by Notke's monumental group of St. George and the Dragon (St. Nicholas' Church, Stockholm; Figure 55), where the dragon's spines are made from real antlers. The group as a whole is, of course, of wood, a medium that could be employed to create intricate, open, thin, and spiky forms impossible in stone.

Bernt Notke of Lübeck

Figure 54: "The Assumption of the Virgin," part of the Altar of the Virgin by Tilman Riemenschneider, 1505–10. In the Herrgottskirche, near Creglingen, Germany.

From Herman Flesche, Gunther and Klaus Beyer, *Tilman Riemenschneider* (1957), published by Veb Verlag der Kunst, Dresden

On the whole, the sculpture produced in France seems to show more decorative restraint. Certainly, the chief French works surviving take the form of large groups, as in the Tonnerre "Entombment" (1450s), or of architectural schemes in which the decoration is clearly subordinate to the figures, as in Châteaudun, Castle Chapel (*c.* 1425).

Restraint is also notable in the chantry chapel of Richard Beauchamp, earl of Warwick (*c.* 1450; Warwick), which has some obvious motifs taken over from the workshop of Sluter. But many of the chantry chapels so common in 15th-century England—for instance, the Henry V Chantry, Westminster Abbey (1440s), or the chantries of John Alcock (*c.* 1488) and Nicolas West (*c.* 1534) at Ely cathedral—show an extraordinary mixture of sculpture and tracery work more reminiscent, as an expression of taste, of Germany or Spain.

The full impression of such profusion can now best be judged from the Chapel of Henry VII (*c.* 1503–*c.* 1515; Westminster Abbey), which is unique in England for the amount of sculpture that has been preserved.

Spanish 15th-century sculpture also tended to be extremely ornate. A number of huge, carved high altarpieces survive—for instance, in the cathedrals of Burgos (1486–88) and Toledo (begun 1498). Some of the altar pieces, like that at Toledo, were designed and executed under the direction of German or Netherlandish artists (Figure 56).

The change from late Gothic to Renaissance was superficially far less cataclysmic than the change from Romanesque to Gothic. In the figurative arts, it was not the great shift from symbolism to realistic representation but a change from one sort of realism to another.

Architecturally, as well, the initial changes involved decorative material. For this reason, the early stages of Renaissance art outside Italy are hard to disentangle from late Gothic. Monuments like the huge Franche-Comté chantry chapel at Brou (1513–32) may have intermittent Italian motifs, but the general effect intended was not very different from that of Henry VII's Chapel at Westminster. The Shrine of St. Sebaldus at Nürnberg (1508–19) has the general shape of a Gothic tomb with canopy, although much of the detail is Italianate. In fact, throughout Europe the "Italian Renaissance" meant, for artists between about 1500 to 1530, the *enjolivement,* or embellishment, of an already rich decorative repertoire with shapes, motifs, and figures adapted from another canon of taste. The history of the northern artistic Renaissance is in part the story of the process by which artists gradually realized that Classicism represented another canon of taste and treated it accordingly.

But it is possible to suggest a more profound character to the change. Late Gothic has a peculiar aura of finality about it. From about 1470 to 1520, one gets the impression that the combination of decorative richness and realistic detail was being worked virtually to death. Classical antiquity at least provided an alternative form of art. It is arguable that change would have come in the north anyway and that adoption of Renaissance forms was a matter of coincidence and convenience. They were there at hand, for experiment.

Refot, Stockholm

Figure 55: "St. George and the Dragon," wood sculpture by Bernt Notke, 1483–89. In St. Nicholas' Church, Stockholm. Approximately 3.05 × 4.19 m.

Figure 56: "The Washing of the Feet," polychromed wood, detail from the retable of the High Altar, 1498. In Toledo cathedral, Spain.
Archivo Mas, Barcelona

Their use was certainly encouraged, however, by the general admiration for Classical antiquity. They had a claim to "rightness" that led ultimately to the abandonment of all Gothic forms as being barbarous. This development belongs to the history of the Italian Renaissance, but the phenomenon emphasizes one aspect of medieval art. Through all the changes of Romanesque and Gothic, no body of critical literature appeared in which people tried to evaluate the art and distinguish old from new, good from bad. The development of such a literature was part of the Renaissance and, as such, was intimately related to the defense of Classical art. This meant that Gothic art was left in an intellectually defenseless state. All the praise went to ancient art, most of the blame to the art of the more recent past. Insofar as Gothic art had no critical literature by which a part of it, at least, could be justified, it was, to that extent, inarticulate. (A.Ma./Ed.)

The "rightness" of Classical antiquity

The Renaissance

ITALY

The revival of Classical learning in Italy, which was so marked a feature of Italian culture during the 15th century, was paralleled by an equal passion for the beauty of Classical design in all the artistic fields; and when this eager delight in the then fresh and sensuous graciousness that is the mark of much Classical work—to the Italians of that time, seemingly the expression of a golden age—became universal, complete domination of the Classical ideal in art was inevitable.

This turning to Classical models was less sudden and revolutionary than it seemed. Throughout the history of Romanesque and Gothic Italian art, the tradition of Classical structure and ornament still remained alive; again and again, in the 12th and 13th centuries Classical forms—the acanthus leaf, moulding ornaments, the treatment of drapery in a relief—are imitated, often with crudeness, to be sure, but with a basic sympathy for the old imperial Roman methods of design. Nicola Pisano, at work in the mid-13th century, was but the first of many Italian artists, particularly sculptors, to turn definitely to Roman antecedents for inspiration.

Early Renaissance. Sculpture was the first of the arts in Florence to develop the Renaissance style. Some would date the beginning of the Renaissance to the sculptural competition in 1401 for the bronze doors of the Baptistery of the cathedral of Florence; others would propose the commission to Donatello and Nanni di Banco in 1408 for four seated saints for the facade of the cathedral. The competition reliefs for the bronze doors, submitted in 1402, reveal a change in attitude toward sculpture, and the figures of the Evangelists are the manifestation of that change. The development of Florentine sculpture roughly parallels the development in painting from a dignified monumental style to a relaxed sweetness, although there is no one in painting to approach the rich inventive genius of Donatello.

Donatello, like his friends the architect Brunelleschi and the painter Masaccio, was one of the most outstandingly original artists in Western history. He undoubtedly was influenced by the concepts of antiquity current in Florence, but there was relatively little antique sculpture visible for him to study in his formative years. He first appears as a mature genius working on two of the major projects of the 15th century, the sculptural decoration of the cathedral of Florence and of the guild church of Or San Michele.

The genius of Donatello

His "St. George," begun *c.* 1415 for the niche of the Armourer's Guild at Or San Michele, indicates the new direction in sculpture (Figure 57). Here he reveals such a deep knowledge of the human figure at rest and in movement that he may already have begun his investigation into proportion and the statics and dynamics of the human figure. But the tension between repose and action—the representation, in fact, of pause—also is a psychological achievement, hardly to be matched in earlier sculpture. It is noteworthy, too, that the monumental simplicity and power of the piece is achieved by such a subtle manipulation of the planes and such a technical virtuosity in

Alinari—Art Resource/EB Inc

Figure 57: "St. George," bronze copy of a marble statue by Donatello, begun *c.* 1415. In Or San Michele, Florence. (The original statue has been transferred to the Bargello, Florence.) Height 2.08 m.

carving the marble that the observer is rarely concerned with the material. The figure is neither flesh nor stone; it simply is.

In the relief under the niche occupied by "St. George," Donatello introduced another great innovation that was to have unlimited repercussions in Florentine art. Relief has always been a problem for sculptors because it must follow a narrow path between the two-dimensionality of painting and the three-dimensionality of full-round sculpture. Donatello conceived of a very low relief in which the subtle modelling of planes suggests the illusion of depth and figures moving in space while still respecting the integrity of the plane. He continued to develop the potentialities of this relief style throughout his long career and strongly determined the kind of relief sculpture executed in Florence.

In his brief career Nanni di Banco was as prolific and inventive as Donatello. In his earliest works, such as the "Isaiah," he approached more closely the Classic ideal than did Donatello, and in his late work at the Porta della Mandorla he began to evolve a relaxed style that was to have its greatest impact after mid-century. About 1411–13 he executed the "Quattro Santi Coronati" ("Four Crowned Saints") for the niche of the woodworkers and stoneworkers guild at Or San Michele. In this commission he solved one of the most difficult problems facing the sculptor, that of the group conceived in the round. Although some of the figures still retain certain Gothicizing elements in the draperies and in the heads, the major impression is of a group of Roman senators born again in the Renaissance. The group is bound together by the spatial relation of one to the other and by a kind of mute conversation in which they are all engaged.

Lorenzo Ghiberti won the competition for the bronze doors of the Baptistery. He began work in 1403 and set the doors in place in 1424. Ghiberti's fame rests upon his second set of doors, the "Gates of Paradise" (1425–52). The gilded bronze reliefs are treated almost like paintings, for they are rectangular in format and contained within a frame. Unlike the earlier doors, in which the ground plane is simply a neutral backdrop, it is here treated in such a way that it suggests sky and space. Figures are placed in landscape or in perspectivally rendered architecture to suggest a greater depth to the relief than actually exists. At the time that he was executing his first set of bronze doors, Ghiberti undertook to cast the first life-sized bronze statue since antiquity, his "St. John the Baptist" (1412–16) for Or San Michele. Although the figure and its draperies reveal Ghiberti's strong adherence to a late Gothic style, with this work he moved technically into the Renaissance. The influence of Donatello and Nanni di Banco liberates the "St. Matthew" of 1419–22, for Or San Michele, from the older traditions. Ghiberti achieved fame in his own time as a bronze founder and as the master of the shop in which many sculptors and painters of the early Renaissance were trained.

Ghiberti's "Gates of Paradise"

The Sienese sculptor Jacopo della Quercia was the most important sculptor of 15th-century Siena. He executed the Fonte Gaia (1414–19), a public fountain for the Piazza del Campo, the main square of Siena, and was awarded the commission for a baptismal font in the baptistery of Siena cathedral. Always a procrastinating artist, he postponed work on the font to such a degree that the reliefs were finally awarded to other sculptors, including Donatello and Ghiberti. Jacopo's major work is the relief sculpture around the main portal of S. Petronio, Bologna (1425–38). The sculptural treatment of the low relief figures and the suggestion of a space adequate to contain them parallels the painting of Masaccio. The dramatic vigour and powerfully conceived forms had a great influence on the young Michelangelo.

Donatello dominated Florentine sculpture of the second quarter of the 15th century. He executed a series of prophets and a "Cantoria," or singing balcony, for the cathedral, saints for Or San Michele, decorative reliefs and bronze doors for the Old Sacristy of S. Lorenzo, and a bronze "David" (now in the Bargello, Florence) that comes closer to recapturing the spirit of antiquity than any other work of the early Renaissance—indeed, the very idea of a free-standing sculpture of a nude hero was without precedent since antiquity. During the decade 1443–53 Donatello was in Padua executing the equestrian statue of Gattamelata to stand in front of the church. Erasmo da Narni, called Gattamelata, was a condottiere, or leader of mercenary troops, who rose to a position of importance. The statue is an idealization of nature in both horse and rider and a reinterpretation of antiquity. Donatello certainly knew the antique statue of Marcus Aurelius in Rome during his stay there (1431–33). He uses the concept of antiquity, the pose of the antique bronze horses at St. Mark's in Venice, and the forms of the war-horse of his own time. The rider is clothed in quasi-antique armour and bears little or no resemblance to the effigy on Gattamelata's tomb inside the church. Donatello is not concerned with particulars but with the idealized and generalized aspects of man that reveal his potential nobility. The "Gattamelata" states the basic concept that almost all equestrian statues have followed since that time. Donatello's presence in Padua gave rise to a productive local school of bronze sculptors and workers, and his reliefs on the high altar there influenced painters and sculptors of northern Italy.

Donatello's "Gattamelata"

One of his first works upon his return to Florence was a wooden statue of Mary Magdalene for the baptistery of the cathedral. The nervous energy and conscious distortion of forms that may be detected in all his work becomes explicit in the emaciated figure clothed in her own hair. This same emotionalism and distortion is even more pronounced in his last work, the pulpits for the church of S. Lorenzo in Florence.

Antonio Pollaiuolo expresses in his sculpture the same sort of muscular activity and linear movement as in his painting—he has the energy but not the interest in emotion found in Donatello. His small bronze "Hercules and Antaeus" (*c.* 1475; Bargello, Florence) is a forceful depiction of the struggle between these two powerful men from classical mythology. The angular contours of the limbs and the jagged voids between the figures are all directed toward expressing tautness and muscular strain, and the work is one of the earliest examples of the statuette in modern times.

The popularity of small bronzes, usually of secular, often of pagan, subjects and sometimes objects of utility (inkwells, candleholders, and so on), increased in popularity toward the end of the century. The elegant, polished antique gods made by Antico in Mantua and the brilliantly modelled satyrs made by Riccio in Padua set a standard in such works that has never been excelled. Bronze statuettes were made by almost all the major sculptors of the 16th century in Italy.

In complete contrast with Pollaiuolo, Desiderio da Settignano is perhaps best known for his portraits of women and children, although he also executed two public monuments of major importance in Florence—the tomb of Carlo Marsuppini in Sta. Croce (*c.* 1453–55) and the "Tabernacle of the Sacrament" in S. Lorenzo (1461). The tabernacle, which was probably assembled and completed by assistants after Desiderio's death, indicates the new trends taking shape in Florentine sculpture. The central panel employs a perspectivized space. The figures moving into that space are defined in a linear manner that emphasizes contours and billowing draperies to suggest movement. The lateral, full round figures of angels are modelled with a delicacy and subtlety of surface to create relaxed and sweet figures very different from Donatello's strong, virile early saints.

The "sweet style" of the Florentine school

Antonio Rossellino collaborated with his older brother Bernardo on the tomb of Leonardo Bruni (*c.* 1445–49) in Sta. Croce but soon became the dominant personality in the family business. The great sculptural complex of the Cardinal of Portugal tomb (1461–66) in S. Miniato al Monte at Florence reveals the same general tendencies as Desiderio's contemporary work. The tomb is decorated with soft and relaxed angels and a tender Madonna and Christ Child in the roundel (Figure 58). Similar tendencies can be found in such artists as Agostino di Duccio, Mino da Fiesole, and Luca della Robbia.

Andrea del Verrocchio was more interested than these sculptors were in movement, which he expressed in a

Figure 58: The Cardinal of Portugal tomb, marble sculptural complex by Antonio Rossellino, 1461–66. In the church of S. Miniato al Monte, Florence.

Alinari—Art Resource/EB Inc

somewhat restrained manner. His group of "Christ and St. Thomas" for Or San Michele (*c.* 1467–83) solves the problem of a crowded niche by placing St. Thomas partly outside the niche and causing him to turn inward toward the figure of Christ. His large equestrian statue of Bartolomeo Colleoni (1483–88) in Venice descends from Donatello's "Gattamelata," but a comparison of the two works reveals Verrocchio's evidence of greater interest in movement. The "Putto with Dolphin" (*c.* 1479; formerly in the Palazzo Vecchio, Florence, but now replaced by a copy) is at once an exquisite fountain decoration, an antique motif restated in Renaissance terms, and the clearest statement of Verrocchio's interest in suggested movement. The child in the piece is seen to be turning; the movement is reinforced by the fish, and the suggestion of motion culminates in the actual movement of the water spouting from the dolphin's mouth. Verrocchio also reveals his indebtedness to Desiderio in the way he treats the surfaces.

High Renaissance and Mannerism. Sixteenth-century sculpture is dominated by the figure of Michelangelo. Although he was born and trained in the 15th century, his style and the bulk of his creations place him firmly in the 16th century. Michelangelo's example was so powerful that Mannerist Florentine artists such as Bartolommeo Ammannati and Baccio Bandinelli could only struggle feebly against it. Others, such as Vincenzo Danti, found it easier to succumb and to follow docilely. Jacopo Sansovino effectively escaped the influence of Michelangelo by transferring his activities to Venice. In Padua a group of bronze workers continued to develop the tradition of fantastic and often beautiful small bronzes that had its origins in Donatello's shop. It was only toward mid-century with artists such as Benvenuto Cellini or at the end of the century with Giambologna that Florentine sculpture found individuals who were able to assimilate Michelangelo's pervasive influence.

Michelangelo Buonarroti is said to have learned sculpture from the minor Florentine sculptor Bertoldo di Giovanni, who provided a link with the tradition of Donatello. An early work, the "Madonna of the Stairs" (*c.* 1492; Casa Buonarroti, Florence), reflects a type of Donatello Madonna and Donatello's very low relief. After the expulsion of the Medici from Florence, Michelangelo fled to Bologna; there he executed three figures for the tomb of S. Domenico and saw the powerful reliefs of Jacopo della Quercia. By 1496 he was in Rome, where he carved a "Bacchus," now in the Bargello, Florence. Michelangelo recaptures the antique treatment of the young male figure by the soft modulation of contours. The figure seems to be slightly off-balance, and the parted lips and hazy eyes suggest that he is under the influence of wine. The little faun also joins in the Bacchic revel by slyly stealing some grapes. In his first major sculptural work the 21-year-old artist succeeded in capturing the spirit of the antique as no artist before him had done. The "Pietà" (today in St. Peter's), commissioned by a French cardinal, was begun immediately upon the completion of the "Bacchus." The motif of the pietà is German in origin, but it is so completely transformed by Michelangelo that the work is one of the harbingers of the High Renaissance. The robes of the Madonna are exaggerated to create a solid base for the pyramidal composition. The figure of Christ is bent and twisted, in part to express the suffering of the crucifixion and in part to make it conform to the contours of the pyramid. All is directed toward creating a calm, dignified, and stable composition that expresses emotion and religious fervour by implication rather than by overstatement. The work is carried to a higher degree of finish than any of the succeeding works, and it is one of the few that Michelangelo signed.

Michelangelo's mastery of the antique

In 1501 Michelangelo was recalled to his native city of Florence to execute an over-life-size figure of "David." When the piece was completed, Michelangelo's contemporaries judged it too important to place out of sight high up on the cathedral, as had been originally proposed, and a committee voted to place it in front of the Palazzo Vecchio, the seat of Florentine civic government. Michelangelo's technical virtuosity is dramatically demonstrated by the fact that he extracted a figure about 14 feet (four metres) tall from a spoiled block. The youthful David was one of the symbols of Florence. Michelangelo sees him as a slightly awkward adolescent with large hands

Michelangelo's "David"

Alinari—Art Resource/EB Inc

Figure 59: "David," marble statue by Michelangelo, 1501–04. In the Accademia, Florence. Height 5.49 m.

and feet, the body of a boy, and the head of a young man—a powerful figure who has not yet realized his full potential. The balance of the figure is subtly arranged to keep the bearing leg under the head while permitting the apparently nonbearing leg to be relaxed. The positions are reversed in the arms, giving the cross-axis balance of working and relaxed members. The head turns to the left to meet Goliath and the stone of the sling is concealed in the right hand. It is this subtle balance and adjustment of parts to create a unified and harmonious whole that places this work firmly in the High Renaissance style that was appearing simultaneously in painting and architecture (Figure 59).

While in Florence from 1501 to 1505, Michelangelo carved "Madonna and Child" for Notre-Dame in Brugge. He began but did not finish a "St. Matthew" for the cathedral, and he painted the "Holy Family" (*c.* 1503–05; Uffizi, Florence), his reply to Leonardo's eminently popular "The Virgin and Child with St. Anne." In competition with Leonardo he began but did not finish the "Battle of Cascina" for the Palazzo Vecchio. On command of Julius II he returned to Rome.

The Roman years (1506–16) are characterized by what Michelangelo later called the tragedy of the tomb. He had been called to Rome to execute a monumental sepulchre for Pope Julius II. The Pope's financial difficulties and the jealousies of the papal court diverted the artist from the tomb to the painting of the Sistine ceiling. The death of Julius in 1513 caused the heirs to press for a smaller tomb and rapid completion. After many years of negotiations, in 1545 a much-reduced version was set in place in S. Pietro in Vincoli, instead of in St. Peter's as originally planned. The figures by Michelangelo for the tomb are now widely scattered. Only the "Moses" remains in place from the original projects. This figure, which recalls so strongly Donatello's "St. John the Evangelist," was intended to be placed well above the observer's head and is so adjusted. The "Dying Slave" and the "Bound Slave" are now in the Louvre. The "Victory," also intended for the tomb, was executed *c.* 1532–34 in Florence, where it has remained. Four unfinished figures of slaves were carved before 1534 and remained in Florence, where they once formed part of the grotto decoration at the Pitti Palace.

With the election of Pope Leo X in 1513, Michelangelo was diverted from his projects and sent to Florence to design a facade for S. Lorenzo, a church under Medici patronage. Although Michelangelo promised that the facade would become the showplace of Italian sculpture, nothing came of the project. He was assigned instead to construct a tomb chapel as a pendant to Brunelleschi's Old Sacristy, and later to provide suitable housing for the Medici library in S. Lorenzo. While engaged in these projects Michelangelo was also put in charge of the fortifications of Florence prior to and during the siege of 1529. He complained, justly, that no one can plan and execute three projects simultaneously.

The Medici tombs (1520–34) gave the artist the opportunity to plan the architectural setting of his sculpture and to control both the light cast on the work and the position of the observer. Since the chapel was originally planned to contain the tombs of the Medici popes Leo X and Clement VII, it is best seen from behind the altar, where the papal celebrant of the mass for the dead would have stood. On the left is the tomb of Giuliano, on the right the tomb of Lorenzo, and before the observer the Madonna and Christ Child with the Medici patron saints, Cosmas and Damian; and beneath the two sarcophagi respectively lie the recumbent figures of "Night" and "Day," and "Dawn" and "Dusk."

The "Pietà," or "Deposition," in the museum of the cathedral of Florence dates from around 1550 and may have been intended by Michelangelo for use in his own tomb. The figure of Nicodemus is a self-portrait and indicates Michelangelo's deep religious convictions and his growing concern with religion. His final work, the "Rondanini Pietà" (1552–64), now in the Castello Sforzesco, Milan, is certainly his most personal and most deeply felt expression in sculpture. The artist had almost completely carved the piece when he changed his mind, returned to the block, and drastically reduced the breadth of the figures. He was working on the stone 10 days before he died, and the piece remains unfinished. In its rough state the "Rondanini Pietà" clearly shows that Michelangelo had turned from the rather muscular figure of Christ of his earlier works (as can be seen from the partially detached original right arm) to a more elongated and more dematerialized form.

Followers of Michelangelo

Whether in Rome or Florence, Michelangelo had a strong influence on sculptors of the 16th century. Vincenzo Danti followed closely in Michelangelo's footsteps. His bronze "Julius III" of 1553–56 in Perugia is derived from Michelangelo's lost bronze statue of Julius II for Bologna. Many of his figures in marble are only free variations on themes by Michelangelo. In much the same way, Baccio Bandinelli attempted to rival the monumentality of Michelangelo's "David" and the complexity of his "Victory" in the statue of "Hercules and Cacus" (1534), which was placed as a companion to the "David" in front of the Palazzo Vecchio. Bartolommeo Ammannati should be best known for his design of the bridge of Sta. Trinità in Florence, but his most visible work is the Neptune Fountain (1560–75) in the Piazza della Signoria, with its gigantic figure of Neptune turned toward the "David" in presumptuous rivalry.

Benvenuto Cellini through his celebrated autobiography has left a fuller account of his picturesque life than that of any other artist of the 16th century. He was in Rome from 1519 to 1540 and was one of the defenders of the pope during the siege of the Castel Sant'Angelo. In France from 1540 to 1545, he executed there the celebrated saltcellar for Francis I and the "Nymph of Fontainebleau" (Louvre). The saltcellar is at once an example of 16th-century conspicuous consumption and of Mannerist conceits in art. It is of solid gold, which is covered in part by enamels as though it were a base metal. It was designed for use as a functional object upon the King's table to hold nothing more than common table salt. On his return to Florence in 1545 Cellini received the commission to cast the bronze "Perseus," now in the Loggia dei Lanzi, Florence, which he describes in some detail in his *Autobiography.* The youthful figure of Perseus seems to retain some of the airiness from his flight on the winged sandals of Hermes. He holds aloft the head of the Medusa in an outstretched arm, thus creating an open composition that exploits to the full the potential of the bronze medium. Void is almost as important as solid in this light and airy composition that would have been unthinkable and impossible in marble. Cellini intended the figure to be seen from a variety of viewing points, a relatively new idea in sculpture of this sort, and he leads the observer around by the position of the arms and the legs.

Giambologna

Florentine sculpture at the end of the 16th century was dominated by the Fleming Giambologna and by his shop assistants. Giambologna went to Italy for study shortly after mid-century and settled in Florence in 1557. His earlier major work in Italy is the Fountain of Neptune (1563–66) in Bologna. By early 1565 he had also cast the earliest of his many versions of the bronze "Flying Mercury" that is his most famous creation. The ideas of Cellini's "Perseus" are here carried to their logical conclusion. The god borne along on the air by his winged sandals touches earth only on the slenderest base possible, which is, in fact, represented as a jet of air from the mouth of a wind god. The statue is perfectly balanced according to principles discovered early in the 15th century, yet the outthrust arms and legs give it a feeling of movement and of lightness. Giambologna understood Michelangelo's *figura serpentinata,* the upward spiralling composition, better than any sculptor of the 16th century. His marble group of the "Rape of the Sabines" (1579–83), in the Loggia dei Lanzi, Florence, interweaves three figures in an upward spiralling composition that prefigures the Baroque. Outside Florence, at the present Villa Demidoff in Pratolino, he carved a figure of the Apennines (1581) that seems to be a part of the living rock; it is an excellent example of late Mannerism, in which a paradoxical relationship between art and nature is often cultivated. As the favourite sculptor of the Medici, Giambologna and his prolific shop

dominated Florentine sculpture at the end of the 16th century, training artists who were to carry late 16th-century ideas into the rest of Europe and prepare the way for the nascent Baroque. (J.R.Sp./Ed.)

Venetian sculpture

In sculpture, Venice was less independent of Florence and Rome than in painting. The major 16th-century impetus came from Jacopo Sansovino, a central Italian who arrived in Venice in 1527. Sansovino never adopted the full-scale Mannerism of Florence, and his style retained a High Renaissance flavour, but his pupils Danese Cattaneo and Alessandro Vittoria were selectively able to develop the more mannered aspects of Sansovino's style into a Venetian species of Mannerism.

Vittoria stands closer to Florentine style than his contemporaries in painting, particularly in his decorative work, and his small bronzes display a serpentine grace surpassed only by Giambologna in Florence. His marble figures are, however, often more directly expressive than those of Florentine sculptors. His altarpiece for S. Francesco della Vigna (1561–63) conforms with the attenuated canons of Mannerist elegance. In sculpture as in painting, the narrative Venetian style proved to be more easily adaptable to the demands of the Counter-Reformation than the abstract artiness of central Italian Mannerism. The work of Vittoria and of the painter with whom he was most closely associated, Palma il Giovane, seems to anticipate many of the characteristics of Baroque art.

MANNERIST SCULPTURE OUTSIDE ITALY

In the north of Europe, Giambologna's influence was paramount. Both Hubert Gerhart and Adriaan de Vries, the leading exponents of northern Mannerist sculpture, can be considered as followers of the expatriate Fleming. Gerhart worked (1583–94) for Hans Fugger at Kirchheim, Augsburg, and at Amsterdam under de Sustris, and for the archduke Maximilian I of Bavaria, at whose court he produced bronze figures of considerable accomplishment (1598–1613). De Vries joined Bartholomaeus Spranger in 1601 at Rudolf's court in Prague. His "Psyche with Three Cupids" (Nationalmuseum, Stockholm) is a characteristic example of his stylishness—a wonderful satin finish,

By courtesy of the Nationalmuseum, Stockholm

Figure 60: "Psyche with Three Cupids," bronze sculpture by Adriaan de Vries, *c.* 1593. In the Nationalmuseum, Stockholm. Height 1.88 m.

Figure 61: Nymphs from the "Fountain of the Innocents," Paris, relief panels (after plaster casts), by Jean Goujon, 1547–49. About life-size.

J.E. Bulloz

spiralling complexity, and a soaring grace reminiscent of Giambologna's "Mercury" (Figure 60).

As in painting, France owed its early acquisition of Mannerist sculptural style to Italian artists at Fontainebleau, to Primaticcio's stucco style, and to Cellini. Jean Goujon began from this point of inspiration, and his decorations for the "Fountain of the Innocents" at the Louvre (1547–49) possess a sophisticated refinement *all'antica* unequalled by any non-Italian artist of the period (Figure 61).

The influence of Primaticcio's suave stucco decorations is even more apparent in the early work of the other great French sculptor of the century, Germain Pilon. This is not surprising since his elegant "Monument for the Heart of Henry II" was probably completed under Primaticcio's supervision. His statues for Primaticcio's Tomb of Henry II, however, show him moving toward greater naturalism and expressiveness. In his later works Pilon achieved a freedom of plasticity and feeling for texture that anticipated Baroque developments.

Spanish Renaissance sculpture at first relied heavily upon visiting Italians, led by Andrea Sansovino, but with the advent of Ordóñez, Diego de Siloé, and the painter-sculptors Machuca and Beruguete, a native Spanish school of Mannerism was formed. Like his father (the painter Pedro), Alonso Beruguete studied in Italy. On his return to Spain about 1517, he began to develop an elaborately pictorial style in sculptural complexes of great originality. The fluid quality of his designs reaches its peak in the surging motions of the "Transfiguration Altar" (1543–48) for Toledo cathedral. Beruguete's greatest successor at Valladolid was Pompeo Leoni, who collaborated with his father, Leone, on portraits of Charles V, composed in a disciplined and sternly Roman style, quite different from the expressive fluency of native Spanish sculpture that reemerged at the turn of the century in the few sculptures of polychromed wood by El Greco. (M.J.Ke.)

The Baroque period

ITALY

Early and High Baroque. At the beginning of the 17th century, sculpture in all of Italy, with the exception of Florence, was at a low ebb; and the dry, frankly propagandist nature of the decoration of the Borghese and Sistine chapels in Sta. Maria Maggiore, Rome, reveals this only

Figure 62: "Ludovica Albertoni," marble effigy by Gian Lorenzo Bernini, *c.* 1674. In the Altieri Chapel, S. Francesco a Ripa, Rome.
Anderson—Alinari from Art Resource/EB Inc

too clearly. With Stefano Maderno and Camillo Mariani a slightly more imaginative interpretation of the demands of the Council of Trent is to be found, while certain aspects of the work of Pietro Bernini (1562–1629) were to have considerable influence on his son Gian Lorenzo. The first breath of the new Baroque spirit, however, is to be found in the immense vitality of the equestrian monuments in Piacenza (1612–25) by Francesco Mochi; and a comparable fiery vigour is the keynote of the fresco "Aurora" by Guercino in the Casino Ludovisi, Rome (1621–23). The forms are pierced and opened up, and the momentary, unstable poses, with draperies fluttering and tails lashing, give a vivid movement that releases the figures from the Mannerist spell.

No field was more congenial to the spirit of Baroque art than sculpture carried out on a conspicuous scale. The Baroque artist achieved dramatic pictorial unity by abolishing the traditional limits separating painting, sculpture, and architecture. The solid masses of sculpture and even of architecture were made to move in space by means of such motive forms as undulations; sculpture was transformed by such painter's devices as richly varied illusionistic textures, coloured materials, and irregularly dappling light effects.

The genius of Bernini

Gian Lorenzo Bernini, the greatest sculptor of the 17th and 18th centuries, established the sculptural principles for those two centuries in a series of youthful works of unrivalled virtuosity, as the "Apollo and Daphne." Stone was now completely emancipated from stoniness by open form and by an astonishing illusion of flesh, hair, cloth, and other textures, pictorial effects that had earlier been attempted only in painting. These qualities made what his contemporaries called his "speaking portraits" seem unprecedentedly alive; portrait sculpture for two centuries was a variation of these innovations. In the statue of St. Longinus in St. Peter's in Rome, Bernini created the characteristic formula of Baroque sculpture by throwing the draperies into a violent turmoil, the complicated and broken involutions of which are not rationally explained by the figure's real bodily movement but seem paroxysmally informed by the miracle itself. The passion with which he imbued his sculptured figures, capturing the most transitory states of mind, reached its apogee in the representation of the ecstasy of St. Teresa in the Cornaro Chapel, Sta. Maria della Vittoria, Rome (1645–52) and in the figure of the expiring Ludovica Albertoni (Figure 62) in the Altieri Chapel, S. Francesco a Ripa, Rome (*c.* 1674). The former is generally considered the masterpiece of Baroque religious sculpture and shows how Bernini could organize the arts of architecture, painting, and sculpture in an overwhelming assault on the senses that dispels the resistance of the intellect. This ambitious plan was typical of the mature Bernini, whose spiritual and artistic aspirations exceeded the scope of his early secular salon statues. His later works were largely religious and unprecedentedly vast in scale, as in the dazzling "Cathedra Petri," which covers the whole end of St. Peter's in Rome with a teeming multitude of figures.

The tombs of Bernini are magnificent spectacles in which symbolic figures, clothed in sweeping draperies, with rhetorical gesture and expressive features, share in some emotional experience, theatrically depicted. An example is the tomb of Alexander VII in St. Peter's, Rome. The pontiff, set in a great apse, kneels on a high pedestal about which Charity, Truth, Justice, and Wisdom weep disconsolately while Death, a skeleton, raises the great draperies of polychrome and gold that veil a darkened doorway. Another work, the fountain of the Triton in the Piazza Barberini, Rome, from which all clarity of profile or of shadow, all definiteness of plane, are removed, is also characteristic of Bernini's style, widely imitated throughout Europe.

Bernini's art was the basis of all Baroque sculpture, but his example was not always followed, and the work of his more restrained contemporaries, such as Alessandro Algardi (relief of "Meeting of Attila and Pope Leo," 1646–53, St. Peter's, Rome) and the Fleming François Duquesnoy, attracted more approval from theorists of art. The

Giraudon—Art Resource/EB Inc

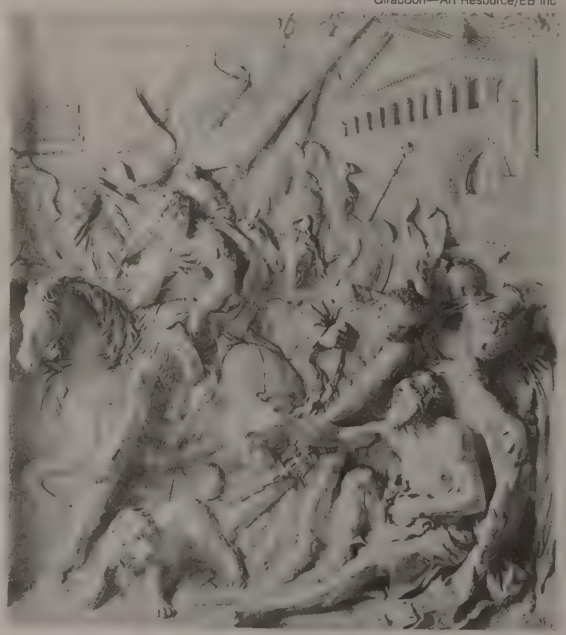


Figure 63: "Alexander and Diogenes," by Pierre Puget, *c.* 1671–93. In the Louvre, Paris.

Figure 64: Fontana di Trevi, Rome, designed and begun by Niccolò Salvi (1732) and completed by Giuseppe Pannini, 1762.
Shostal Inc —EB Inc

latter's "St. Susanna" in Sta. Maria di Loreto in Rome, a figure after the antique but enlivened with Berninian textures, was originally made to look toward the observer and, with a gesture, to direct his attention to the altar. The distinction between art and life that the Mannerists had cultivated was banished by this active participation of the statue in the viewer's space and activities, another important innovation of Bernini.

Late Baroque. In late 17th-century painting, composition became increasingly decorative rather than structural, and there was a loosening of design in the individual figures as well. This dissolution is also to be found in sculpture of the period, such as in the proto-Rococo figures of Filippo Carcani (active 1670–90) in Rome and, to a lesser extent, in those of Filippo Parodi (1630–1702) in Genoa, Venice, and Naples. Outside Venice and Sicily the true Rococo made little headway in Italy.

A more or less classical late Baroque style, best exemplified by the heroic works of Camillo Rusconi in Rome, was dominant in central Italy through the middle of the 18th century. Rusconi's work had considerable influence outside Italy as well.

The latter half of the century saw the emergence of a much lighter and more theatrical manner in the works of Agostino Cornacchini and of Pietro Bracci, whose allegorical figure "Ocean" on the Fontana di Trevi by Niccolò Salvi (completed 1762) is almost a parody of Bernini's sculpture. Filippo della Valle worked in a classicizing style of almost French sensibility, but the majority of Italian sculpture of the mid-18th century became increasingly picturesque with a strong tendency toward technical virtuosity. Complex sculptured groups designed by Luigi Vanvitelli for the park of the palace at Caserta (*c.* 1770) are almost *tableaux vivants* ("living pictures") in a landscape setting, while the Cappella Sansevero de' Sangri in nearby Naples (decorated 1749–66) is one of the most important sculptured complexes of the time. Allegorical groups by Antonio Corradini and Francesco Queirolo vie with each other in virtuosity and include such conceits as fishnets cut from solid marble and the all-revealing shrouds developed by Giuseppe Sammartino. Florentine sculpture of the 18th century is less spectacular, and Giovanni Battista Foggini took back from Rome the compromise style of Ferrarza, while Massimiliano Soldani-Benzi seems to have been instrumental in the brilliant revival there of small-scale bronze statuettes. Giovanni Marchiori worked in Venice with an attractive painterly style, in part based on the wood carvings of Andrea Brustolon; and Giovanni

Archivo Mas, Barcelona

Figure 65: "Pieta," polychromed wood sculpture by Gregorio Hernández, 1617. In the Museo Nacional de Esculturas, Valladolid, Spain. Height 1.8 m.

Maria Morlaiter ran the full gamut to a late 18th-century classicism close to the early works of the great Neoclassical sculptor Antonio Canova.

BAROQUE AND ROCOCO OUTSIDE ITALY

Realism and emotionalism of religious sculpture

Spain. Spanish sculpture of the 17th and 18th centuries exhibits a greater continuity with late Gothic art than does the painting; and the Counter-Reformation demands for realism and an emotional stimulus to piety led to sculpture with glass eyes, human hair, and even real fabric costumes. Italian Renaissance sculpture had made a very limited impact in Spain, and with few exceptions this was in the court ambience only, while Spanish Baroque sculpture is almost entirely religious and of a fundamentally popular nature. Gregorio Hernández in sculptures like the "Pieta" (1617; Museo Nacional de Esculturas, Valladolid, Spain) revealed an emotional realism more Gothic than Baroque (Figure 65); but in the figures of Manuel Pereira there is a clear-cut monumentality and intense concentration comparable to that of Zurbarán. Both were active in Castile, though the main centre of sculptural activity was Seville and Granada, with Juan Martínez Montañés as the dominant personality. The intense realism and deep spirituality of his figures were followed by his pupil Alonso Cano; but in the figures of Cano's pupil Pedro de Mena, his simple monumentality is replaced by a more picturesque and theatrical gracefulness. José de Mora, also a pupil of Cano, took this process even further. But in general the 18th century saw a sad decline in Spanish sculpture.

Flanders. In comparison with painting, the sculpture of the 17th century in the southern provinces is extremely disappointing. The Flemish sculptor François Duquesnoy spent almost all of his career in Rome, while those who remained in Flanders, such as his brother Hieronymus Duquesnoy the Younger, were mostly secondary artists influenced by Rubens. Artus Quellinus the Elder reveals a much more individual style, particularly in his decorations for the Town Hall in Amsterdam, and the tendency toward a painterly style is more pronounced in the work of his son Artus Quellinus the Younger, Rombout Verhulst, and Lucas Faydherbe.

Decline of Antwerp in the late 17th century

The end of the Twelve Years' Truce in 1621 had brought back Antwerp's old troubles, and the control of the Scheldt by the United Provinces was confirmed by the Peace of Westphalia (1648). Economic depression and French aggression in the second half of the 17th century combined to make the southern provinces increasingly provincial, while under the provisions of the Treaty of Utrecht (1713) and the Treaty of Rastatt (1714) the territories passed to Austria. Eighteenth-century painting and sculpture became increasingly weak and provincial, though fantastic pulpits carved by Hendrik Frans Verbruggen, Michel Vervoort, and Theodor Verhaegen provide a remarkable parallel to those in central Europe.

France. Duquesnoy was much admired in France, where the sculptors of Louis XIV (the "Sun King"), such as François Girardon, continued his tradition of setting correct and charming allusions to the antique in a pictorial and spatial context that is wholly Baroque. Girardon's tomb of the Cardinal de Richelieu, in the church of the Sorbonne, Paris, is illustrative of the Baroque monuments of France, calmer and more conservative than those of Italy. The dying cardinal, lying on his sarcophagus and originally gesturing in supplication toward the altar, is upheld by Religion and mourned by Science. The three figures, united by the lines of skillfully arranged draperies, are informed by a solemn and touching sentiment. The academic discipline imposed by the Sun King's ministers, especially Colbert, discouraged less tractable spirits, such as the passionate genius Pierre Puget. His unique expressions of anguish are couched in the physical terms of highly original works like the "Milo of Crotona"; here the composition of a figure rigid with pain is given an almost unbearable tension.

Antoine Coysevox, another of the sculptors of Louis XIV, had begun in the official "academic Baroque" style, but his later works, undertaken after the death of Colbert, are witnesses of the gradual acceptance of the Baroque in France, which now acquired the artistic leadership that Italy had long held over the rest of Europe. At the same time, the style was made lighter, gayer, and more ornamental, in accordance with 18th-century taste, as seen in the famous "Chevaux de Marly" by Guillaume Coustou now marking the entrance to the Champs-Élysées in Paris but designed for Marly, as part of the most innovative outdoor display of sculpture since the 16th-century gardens of Italy. Coustou's bust of his brother Nicolas has a characteristic freshness and informality whereby 18th-century artists avoided the grandeur they found pompous in the Berninian tradition.

This 18th-century style that reduced the Baroque to exquisite refinement was the art of the aristocratic salon and boudoir. The little marble "Mercure" (1741) of Jean-Baptiste Pigalle is almost wholly Berninian, except in its intimacy and deliberate unpretentiousness; even in Pigalle's most ambitious undertakings, the relative scale of the figures is much reduced and the whole composition opened up, in contrast to Bernini's tombs. Nevertheless, the narrative and indeed the allegory of his masterpiece, the tomb of the Maréchal de Saxe (1753; Saint-Thomas, Strasbourg), is as enthralling and memorable as any 17th-century sculpture, although the theme, significantly, no longer seems to be inspired by the Christian faith. At the same time, the more classical current of French sculpture continued and gained importance as the 18th century advanced. The clarified form and continuous, unbroken contours of Étienne-Maurice Falconet's marble "Bather" (1757) adapt the Classic tradition to a pretty and intimate Rococo ideal that is the quintessence of 18th-century taste.

Jean-Antoine Houdon

This Classicism was purified by Jean-Antoine Houdon, who avoided the playful air of the Rococo boudoir in his "Diana" (*c.* 1777) and his marble nude in the Metropolitan Museum of Art, New York City (1782). His portrait sculptures are the ultimate in the 18th-century refinement of Bernini's tradition.

In the context of the rather restrained French sculpture of the 18th century, the blatant sensuality of Clodion (by-name of Claude Michel) is the exception rather than the rule (Figure 66). Portrait busts by Jean-Baptiste Lemoyne and Pigalle follow the direction taken by Coysevox in his "Robert de Cotte," but Augustin Pajou and Houdon

By courtesy of the Metropolitan Museum of Art, New York, bequest of Benjamin Altman, 1913

Figure 66: "Satyr and Bacchante," terra-cotta sculpture by Clodion, *c.* 1775. In the Metropolitan Museum of Art, New York City. Height 58.4 cm.

soon abandoned the Rococo in favour of a Neoclassical approach. Edme Bouchardon, however, flirted only briefly with the Rococo and otherwise remained firmly attached to the classicizing tradition of French sculpture.

England. English sculpture of the early 17th century was very provincial, with Nicholas Stone and Edward Marshall the only English-born sculptors to rise above the general level of mediocrity. Their styles were based on contemporary Netherlandish sculpture with small admixtures of Italian influence; and after 1660 the uncomprehending borrowings of John Bushnell from Bernini serve only to make his figures look ludicrous. The most distinguished English-born sculptor of the second half of the 17th century was Edward Pierce, in whose rare busts is to be found something of Bernini's vigour and intensity. But the general run of English sculpture as represented by Francis Bird, Edward Stanton, and even the internationally renowned woodcarver Grinling Gibbons remained unexceptional. It was not until John Michael Rysbrack from Antwerp settled in England in *c.* 1720, followed by the Frenchman Louis-François Roubillac in *c.* 1732, that two sculptors of European stature were active in England. The busts and tombs of Rysbrack and Roubillac have a power and vitality previously unknown in English sculpture; they were responsible for the revival that took place in the 18th century.

Central Europe. While the influence of Giambologna persisted in some quarters, Hans Krumper and Hans Reichle produced bronze figures less indebted to the Classical tradition but with stronger individuality. Jörg Zürn, whose finest wood carvings are to be seen at Überlingen, and Ludwig Münsterman, in Oldenburg, continued in the Mannerist style, whereas Georg Petel, who came under the influence of Rubens, is almost the only sculptor to reveal the impact of the Baroque. Petel's importance lies mainly in his ivories, and Leonard Kern in Franconia developed a similar Rubensian style for his small statuettes.

Influence of Bernini

Painting and sculpture recovered slowly from the ravages of the Thirty Years' War, and some of the earliest reflections of the high Baroque of Bernini are to be found in the sculpture of Matthias Rauchmiller at Trier (1675) and Legnica (Liegnitz) in Silesia (1677).

Among sculptors in Austria the forces of Classicism were stronger; and the weak north Italian late Baroque styles of Giovanni Giuliani and Lorenzo Mattielli were supplanted by the cool elegance and classical refinement of Georg Raphael Donner. His preference for the soft sheen of lead gave Austrian Baroque sculpture one of its most distinctive features.

Figure 67: Altar of the monastery church at Rohr, Germany, by Cosmas Damian Asam and Egid Quirin Asam, 1718–25.

During the first four decades of the 18th century, Bohemian Baroque art developed almost independently of Vienna. The brilliant rugged stone sculptures of Matyás Bernard Braun and Ferdinand Maximilián Brokoff, with their dynamism and expressive gestures, were truly Bohemian in spirit.

Bavarian Baroque art in the hands of the brothers Egid Quirin Asam and Cosmas Damian Asam was almost entirely confined to churches, and their brilliant development of the theatrical illusionism of Bernini is achieved in the high altar (Figure 67) of the monastery church at Rohr, in Germany (1718–25), and in St. John Nepomuk in Munich (begun 1733). Cosmas Damian's style as a painter was influenced by Rottmayr as well as by the Italian masters whom he studied during his stay in Italy (1711–14), while the sculptural style of Egid Quirin was formed on the south German tradition of wood carving, as well as on Bernini.

The Saxon school

In Upper Saxony there was also a native tradition before the arrival of Permoser, represented by the heavy figures of Georg Heermann and Konrad Max Süssner, both of whom had been active in Prague in the 1680s. Balthasar Permoser was trained in Florence under Foggini, whence he was summoned to Dresden in 1689. His painterly conception of sculpture, derived from Bernini, is revealed in the complex "Apotheosis of Prince Eugene" (1721; Österreichische Galerie, Vienna) and above all in the sculptural decoration of the Zwinger in Dresden initiated during the second decade. Paul Egell was a pupil of Permoser in Dresden at the time of the Zwinger decorations, and in 1721 he was appointed court sculptor at Mannheim. Egell's elongated and refined Baroque figures were an effective counter to the Classicism of Donner, and his personality was decisive in Franconia and the Palatinate during the first half of the century.

Berlin under the Great Elector of Brandenburg had become an increasingly important centre, both politically and artistically; and the full-bodied Baroque style of Andreas Schlüter, as revealed by his equestrian monument to the Great Elector (1696–1708), now at Charlottenburg, was fully in sympathy with the time.

The Rococo style in Bavaria

No hard and fast division can be made between the Baroque and the Rococo in central and eastern Europe, either chronologically or stylistically. The first Rococo decorative ensembles in Germany, the Reiche Zimmer of the Residenz in Munich, were built by the Frenchman François de Cuvilliés in 1730–37, but in painting and sculpture the situation is more complicated. Ignaz Günther, the greatest south German sculptor of the 18th century, was trained under Johann Baptist Straub; the elongated forms of Egell's sculpture at Mannheim, however, deeply impressed him, and his development was toward an almost Mannerist grace and refinement. Günther was capable of the most extraordinarily sensitive characterization of surfaces, even when painted white; and this he combined with an interpretation of character comparable to the late Gothic sculptors, thus giving his figures a realism and immediacy that is almost uncanny. Apart from their lightness and vivacity, however, it is the figures' relationship to the altars on which they are placed that reveals their Rococo quality. Gone are the great coordinated ensembles of the Asams, and instead each figure has a totally separate existence of its own and a balance is only to be found when the church interior is taken as a whole.

Johann Michael Feichtmayr

Swabian sculpture of the period is characterized by the extremely successful partnerships between the sculptors and stucco artists. For Zwiefalten and Ottobeuren Joseph Christian provided the models from which Johann Michael Feichtmayr created the superb series of larger than life-size saints and angels that are the glory of these Rococo interiors. Feichtmayr was a member of the group of families from Wessobrunn in southern Bavaria that specialized in stucco work and produced a long series

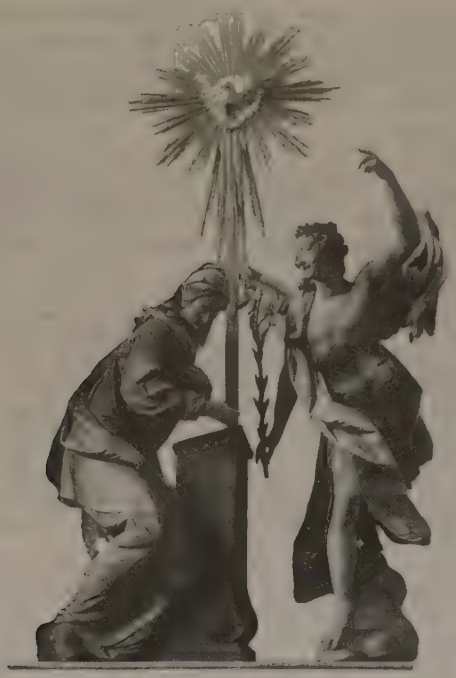


Figure 68: "The Annunciation," painted wood sculpture by Ignaz Günther, 1764. In the abbey church at Weyarn, Bavaria.

Bildarchiv Foto Marburg/Art Resource, NY

of masters, including Johann Georg Übelherr and Joseph Anton Feuchtmayer, whose masterpieces are the Rococo figures at Birnau on Lake Constance. The sculptor Christian Wenzinger worked at Freiburg im Breisgau in relative isolation, but his softly modelled figures have a delicacy that recalls the paintings of Boucher.

Until his death Johann Wolfgang van der Auvera was the most powerful personality in the field of sculpture in the area, but later Ferdinand Dietz at Bamberg pursued an increasingly individual Rococo style that often parodied the growing taste for Neoclassicism. Prussian Rococo sculpture was less distinguished, though the decorations of Johann August Nahl are among the most imaginative in Germany.

Austrian sculpture of the later 18th century, as represented by Balthasar Ferdinand Moll, inclined more toward a realistic Rococo style than to the Classicism of Donner; and, although the strange, neurotic genius Franz Xavier Messerschmidt began in this style, at the end of his career he produced a startling series of grimacing heads when he lived as a recluse in Bratislava.

Russia. The Baroque style as it was imported to Russia from western Europe by the imperial court never amounted to what might properly be termed a Russian Baroque period. A great influx of Western influence during this period, especially under the sponsorship of Peter the Great, did, however, dispel the predominance of Byzantine ideas and forms. The brilliant Baroque busts of Bartolomeo Carlo Rastrelli the Younger established during the early 18th century a distinguished tradition of Russian portrait sculpture that was maintained by Fedot Shubin. The parks and gardens of the Rococo palaces of the empress Elizabeth were adorned with sculpture, but the work was done almost exclusively by Italians and Frenchmen commissioned for the task.

Latin America. With the coming of Europeans to Central and South America, indigenous symbolism and sculptural forms blended with Renaissance realism, Baroque elegance, and subsequent stylistic currents. Indian traits appeared in such European-introduced sculptural forms as the stone crosses that were erected in churchyards; statues, whether by European sculpture or aboriginal pupils, depicted Jesus, the Virgin Mary, saints, and occasionally an earthly benefactor of the church. Materials were of wood, plant fibre pulp coated with canvas and gesso, or plaster. The statues often had real costumes and hair, glass eyes and teeth, and extremely realistic flesh—bloody, bruised, and torn—with taut muscles and distended veins. Gold halos or crowns were added and costume textures were imitated by the gold-leaf-and-paint estofado technique. Many of these were undoubtedly inspired by paintings brought from Europe.

Few sculptors are known by name from the colonial period and fewer attributions are possible. At least a dozen individuals can be identified in Mexico in the 16th century, however, and twice that number in the 17th; the best known are José Cora of Puebla and his nephew Zacarías, and Gudiño of Querétaro. Many were both sculptors and architects, a necessity of the times. In the 18th century considerable artistic stimulus was provided by the Spanish-born Neoclassicist Manuel Tolsa, first director of the Academy in Mexico City, first to produce an equestrian statue in the New World (of Charles IV), and teacher of many sculptors of subsequent fame. The second most important artistic centre of the colonial era was Quito, Ecuador, which was known particularly for its decorative sculpture.

The sculpture is marginally less provincial than the paintings, and, for example, the choir stalls carved by Pedro de Noguera and his assistants for Lima cathedral (1624–26) are of distinguished quality. The Baroque tradition tended to last until well into the 19th century in sculptures such as the robust figures of António Francisco Lisboa (*e.g.*, "O Aleijadinho," or "The Little Cripple"), the greatest sculptor that Brazil has produced.

(P.C.-B./J.Hud./J.Hm./A.Vo./C.I.C./Ed.)

Neoclassical and Romantic sculpture

NEOCLASSICISM

The 18th-century arts movement known as Neoclassicism represents both a reaction against the last phase of the Baroque and, perhaps more importantly, a reflection of the burgeoning scientific interest in classical antiquity. Archaeological investigations of the classical Mediterranean world offered to the 18th-century cognoscenti compelling witness to the order and serenity of Classical art and provided a fitting backdrop to the Enlightenment and the Age of Reason. Newly discovered antique forms and themes were quick to find new expression.

The successful excavations contributed to the rapid growth of collections of antique sculptures. Foreign visitors to Italy exported countless marbles to all parts of Europe or employed agents to build up their collections. The accessibility of the sculpture of antiquity, in museums and private houses and also through engravings and plaster casts, had a far-reaching formative influence on 18th-century painting and sculpture. The great majority of ancient sculptures collected were Roman, although many of them were copied from Greek originals and were believed to be Greek.

In the writing of Johann Joachim Winckelmann, the great German historian of ancient art, Greek art had been considered immeasurably superior to Roman. It is curious, however, how little positive influence the marbles that Lord Elgin took to England from the Parthenon in Athens had on sculpture in western Europe, although they had a great influence on scholars. The ideals of Neoclassical sculpture—its emphasis on clarity of contour, on the plain ground, on not rivalling painting either in the imitation of aerial or linear perspective in relief or of flying hair and fluttering drapery in freestanding figures—were chiefly inspired by theory and by Roman neo-Attic works, or indeed by Roman pseudo-Archaic art. The latter class of art exerted an influence on John Flaxman, who was enormously admired for the severe style of his engravings and relief carvings.

"Decorum" and idealization. Academic theorists, especially those of France and Italy during the 17th century, argued that the costume, details, and setting of a work be as accurate as possible when representing a period and place in the historical past. The 18th century and, in particular, the Neoclassicists inherited this theory of "decorum" and, enabled by all the newly available archae-

Figure 69: "Paolina Borghese as Venus Victrix," marble sculpture by Antonio Canova, 1805–07. In the Borghese Gallery, Rome. 1.60 m × 2.00 m.
Alinari—Art Resource/EB Inc

ological evidence, implemented it more fully than had any of their precursors.

Question of the use of Classical or contemporary dress

A series of monuments to 18th- and early 19th-century generals and admirals of the Napoleonic Wars in St. Paul's Cathedral and Westminster Abbey demonstrate an important Neoclassical problem: whether a hero or famous person should be portrayed in Classical or contemporary costume. Many sculptors varied between showing the figures in uniform and showing them completely naked. The concept of the modern hero in antique dress belongs to the tradition of academic theory, exemplified by the English painter Sir Joshua Reynolds in one of his Royal Academy *Discourses:* "The desire for transmitting to posterity the shape of modern dress must be acknowledged to be purchased at a prodigious price, even the price of everything that is valuable in art." Even the living hero could be idealized completely naked, as in two colossal standing figures of Napoleon (1808–11; Apsley House, London, and Brera, Milan) by the Italian sculptor Antonio Canova. One of the most famous of Neoclassical sculptures is Canova's "Paolina Borghese as Venus Victrix" (1805–07; Borghese Gallery, Rome). She is shown naked, lightly draped, and reclining sensuously on a couch, both a charming contemporary portrait and an idealized antique Venus (Figure 69).

By courtesy of the Thorvaldsen Museum, Copenhagen

Figure 70: "Christ" from "Christ, John the Baptist and the Apostles," marble statue by Bertel Thorvaldsen, 1821. In the Church of Our Lady, Copenhagen. Height 3.36 m.

Relation to the Baroque and the Rococo. Classical academic theories circulating in the Renaissance, and especially in the 17th century, favoured the antique and those artists who followed in this tradition. The artists praised included Raphael, Michelangelo, Giulio Romano, and Annibale Carracci. The slightly later generation of writers added the name of the French painter Nicolas Poussin to the list. The exuberance and "fury" of the Baroque must be avoided, it was argued, because they led to "barbarous" and "wicked" works. Continuing in this tradition, Winckelmann, for example, argued that the Italian Baroque sculptor and architect Bernini had been "misled" by following nature.

Anti-Baroque feeling

Such hostility to Baroque works, however, did not immediately eradicate their influence on 18th-century artists, as can be seen, for example, in an early work by Canova, "Daedalus and Icarus" (1779; Museo Civico Correr, Venice), executed before he had been to Rome. In Canova's tomb of Pope Clement XIV (1784–87; SS. Apostoli, Rome), the Pope, seated on a throne above a sarcophagus, is treated in a dramatically realistic style with hand raised in a forceful gesture reminiscent of papal tombs of the 17th century.

Rococo influence

Although the Neoclassical artists and writers expressed contempt for what they regarded as the frivolous aspect of the Rococo, there is a strong influence of French Rococo on the early style of some of the Neoclassical sculptors. Étienne-Maurice Falconet, Flaxman, and Canova all started to carve and model with Rococo tendencies, which were then gradually transformed into more Classical elements.

Hostile critics of Neoclassical sculpture have tended to compare such works to "a valley of dry bones." Some artists misunderstood the advocacy of Winckelmann and his school to imitate ancient art. Winckelmann meant—as did 17th-century theorists before him, and writers such as Shaftesbury and Jonathan Richardson, who influenced him considerably—imitation to be a means of discovering ideal beauty and conveying the spirit of the original. He did not advocate servile copying of the antique. Unfortunately, spiritless copies were made, and these led to a proliferation of the Neoclassical style and its classification as "frigid." In sculpture some of the important commissions

Figure 71: "Fury of Athamas," marble sculpture by John Flaxman, 1790–92. In Ickworth, Suffolk.
A.F. Kersting

regrettably resulted in this lifeless concept of Neoclassicism. Among the examples are large marbles of Christ (Figure 70), John the Baptist, and the Apostles by the Danish sculptor Bertel Thorvaldsen in the Church of Our Lady, Copenhagen (1821–27 and 1842). Thorvaldsen's marbles, unlike Canova's, lose little when seen only as plaster casts, and indeed the surface of the sculpture was deliberately left neutral and the act of carving left to others.

Classical serenity and restraint

Gestures and emotions in Neoclassical works are usually restrained. In bacchanalian scenes the gaiety is held in check, never bursting into exuberance. In a tragic scene, Andromache does not shed a tear as she mourns the death of Hector. When Flaxman did attempt terror, as in the marble "Fury of Athamas" (1790–92; Ickworth, Suffolk), the violence is forced and unconvincing (Figure 71). Indeed, there hardly exists in any Neoclassical sculptor's work a convincing image of rage. The concept of antique calmness permeated European art. Canova with his "Hercules and Lichas" (1796; Galleria Nazionale d'Arte Moderna, Rome) produced a large marble of exaggerated expression beyond his normal range and to some extent beyond his abilities. Like Flaxman, he was far more successful when carving images of delicacy, bordering on charm.

Prominent early British Neoclassicist sculptors included John Wilton, Joseph Nollekens, John Bacon the Elder, John Deare, and Christopher Hewetson, the last two working mostly in Rome. The leading artist of the younger generation was John Flaxman, professor of sculpture at the Royal Academy and one of the few British artists of the period with an international reputation. The last generation of Neoclassicists included the sculptors Sir Richard Westmacott, John Bacon the Younger, Sir Francis Chantrey, Edward Hodges Baily, John Gibson, and William Behnes.

While Neoclassicism in France was dominated by painting and architecture, the movement did find a number of notable exponents in sculpture. These included Claude Michel, called Clodion, creator of many small Classical figures, especially nymphs; Augustin Pajou; and Pierre Julien. Pigalle's pupil Jean-Antoine Houdon was the most famous 18th-century French sculptor, producing many Classical figures and contemporary portraits in the manner of antique busts. Other contemporary sculptors included Louis-Simon Boizot and Étienne-Maurice Falconet, who was director of sculpture at the Sèvres factory. The slightly younger generation included the sculptors Joseph Chinard, Joseph-Charles Marin, Antoine-Denis Chaudet, and Baron François-Joseph Bosio. The early sculpture of Ingres's well-known contemporary François Rude was Neoclassical.

Important among central European sculptors early in the period was Johann Heinrich von Dannecker. Subsequent Neoclassicists included Johann Gottfried Schadow, who was also a painter but is better known as a sculptor; his pupil, the sculptor Christian Friedrich Tieck; the painter and sculptor Martin von Wagner; and the sculptor Christian Daniel Rauch.

Canova

The most important Italian Neoclassicist was Antonio Canova, the leading sculptor, indeed by far the most famous artist of any sort, in Europe by the end of the 18th century. Canova's position in the following 20 years may be compared only with that enjoyed by Bernini in the 17th century. The differences between their careers, however, are of great importance. Only at the commencement of his career did Bernini carve gallery sculpture for princely collectors, but the majority of Canova's works belong to this category. Both artists remained resident for most of their life in Rome, but whereas Bernini was controlled by the popes and only rarely permitted to work for foreign potentates, Canova's principal patrons were foreigners, and he supplied sculpture to all the courts of Europe. A fine sculptor of varying styles, including austere, sentimental, and horrific, Canova produced an extensive body of work that includes Classical groups and friezes, tombs, and portraits, many in antique dress. He was also a painter, regrettably bad. His pupil and collaborator, Antonio d'Este, is one of the more interesting of the lesser Italian Neoclassical sculptors. Other Neoclassical sculptors in Rome included Giuseppe Angelini, best known for the tomb of the etcher and architect Giambattista Piranesi in the church of Sta. Maria del Priorato, Rome.

In Milan, Camillo Pacetti directed the sculptural decoration of the Arco della Pace. The work of Gaetano Monti, born in Ravenna, can be seen in many northern Italian churches. The Tuscan sculptor Lorenzo Bartolini executed some important Napoleonic commissions. The "Charity" (Pitti Palace, Florence) is one of the more famous examples of his later Neoclassicism. It should be noted, however, that he did not see himself as a Neoclassical artist and that he challenged the idealism that was favoured by Canova and his followers.

The Swede Johan Tobias Sergel, court sculptor to the Swedish king Gustav III, and the Dane Bertel Thorvaldsen, who lived most of his life in Rome, were among the best known Neoclassical sculptors in Europe. Thorvaldsen was the chief rival to Canova and eventually replaced him in critical favour. His work was more severe, sometimes even archaizing in character, and his religious sculpture, most notably his great figure of Christ in the Church of Our Lady in Copenhagen, exhibits a deliberately chilling, sublime style that still awaits sympathetic reassessment. Among his more notable pupils was the Swedish sculptor Johan Byström.

The principal Neoclassicists in Spain were the painter José de Madrazo y Agudo and the sculptor José Alvarez de Pereira y Cubero.

Both leading Russian Neoclassicists were sculptors. Ivan Petrovich Martos studied under Mengs, Thorvaldsen, and Batoni in Rome and became a director of the St. Petersburg Academy. His best works are tombs. Mikhail Kozlovskij contributed to the decoration of the throne room at Pavlovsk.

Apart from the painter Benjamin West, who worked almost entirely in London, the leading Neoclassicists among American artists were sculptors. William Rush produced standing Classical figures including those formerly decorating a waterworks in Philadelphia (now in the Pennsylvania Academy of Fine Arts). In the middle years of the 19th century there came into prominence four sculptors: Horatio Greenough, who executed several government commissions in Washington, D.C.; Hiram Powers, known particularly for his portrait busts; Thomas Crawford, who did monumental sculpture; and William Wetmore Story,

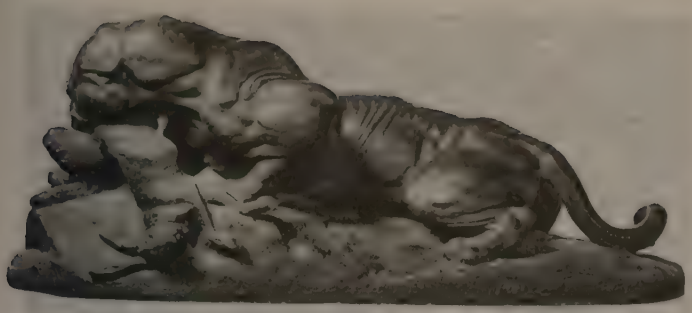


Figure 72: "Jaguar Devouring a Hare," bronze sculpture by Antoine-Louis Barye, 1850–51. In the Louvre, Paris. 41.9 cm × 95.2 cm.

Alinari—Art Resource/EB Inc

who lived and worked in Rome, where he was associated with several other prominent 19th-century Americans.

(D.I.)

19TH-CENTURY SCULPTURE

In the 19th century sculptors throughout the Western world were affected in an unprecedented way by the great public annual exhibitions organized by the Academies. Great patrons at court or among the nobility could still play a very important part in making an artist's reputation, but publicity from these exhibitions was crucial. Among examples of sculptures that attracted sensational publicity of this sort are François Rude's "Neapolitan Fisherboy" (1834; Louvre), Hiram Powers' "Greek Slave" (1843), Auguste Clésinger's "Woman Bitten by a Snake" (1847; Louvre), and Randolf Rogers' "Nydia the Blind Girl" (1858).

In all these sculptures except the last the subject is more or less nude. In all except the first there is a strong narrative interest. In these respects they resemble the prize pieces set by the French Royal Academy of Painting and Sculpture and by its numerous imitators. Unlike those prize pieces, however, these works drew for their subjects not upon Greek or Roman mythology or history: Nydia is a Roman girl but taken from a modern novel about Pompeii, and the Greek slave is a contemporary Christian girl taken captive by the Turks. The old clichés about "academic" sculpture in the 19th century are hopelessly inadequate. The Academies in their educational program often encouraged a heroic but restrained Neoclassicism—their exhibitions, on the other hand, encouraged an appeal to novelty, to sentiment, and to sensationalism (often of an unfortunate kind), involving subjects from modern life and modern literature.

The exhibition piece was often a plaster cast of the original clay model. Several versions in marble or bronze were then made if there was the demand. These would be acquired for the sculpture galleries, conservatories, or gardens of great collectors, as well as for museums, which, for the first time, included collections of modern art. In reduced form they might also make an appearance amid the crowded furnishings of fashionable drawing rooms. Upon the chimneypiece perhaps some miniature scene of jungle violence modelled by Barye and cast in bronze might be displayed, while behind the ferns a marble nude would shrink in vain from male scrutiny.

The proliferation of domestic sculpture was made possible by a series of technical innovations chiefly associated with Paris. Improved reducing machines greatly facilitated the half-size replication of exhibition pieces, and the reproduction of such works on a still smaller scale as bronze statuettes; new methods of sand-casting meant that these bronzes were also available in larger editions and at a lower cost. The reproduction of terra-cotta sculpture also thrived in Paris as it had done in the late 18th century; busts of men of letters and women of fashion, together with groups of seductive nymphs, were always the most popular subjects. The miniature sculptures (often also reproductions of larger works) in biscuit porcelain, which had also been produced in 18th-century Paris, also continued to be popular in England for a while, as well as France.

Exalted notions of the artist's role, inculcated by the Academies and dramatized by Romantic literature, did little to encourage sculptors to involve themselves with what was often described as "mere" ornament. Mechanical methods—more and more sophisticated machinery for turning and pointing, as well as reducing machinery and novel techniques of casting—were often employed with great success. This resulted, however, not only in more bad sculpture than before but also in more badly carved and cast ornament in architecture, furniture, and metalwork. In Paris, however, the fertile genius of Albert Carrier-Belleuse particularly excelled in devising such objects as gasoliers supported by pretty girls in a luxurious style that combined elements from the art of the 16th, 17th, and 18th centuries. In England, Alfred Stevens, inspired by the versatility of the Italian Renaissance, was happy to devote himself to the design of cutlery and fire grates, and, at the end of the century, Alfred Gilbert, creator of the most remarkable metropolitan fountain since the Renaissance (the Eros in Piccadilly Circus), also became the first sculptor of the foremost rank since Cellini to devote himself wholeheartedly to the art of the goldsmith.

Perhaps the least successful aspect of 19th-century sculpture was the large-scale relief panels and pedimental ornaments and niche stances on churches and public buildings—the individual styles encouraged by the exhibition were inappropriate, and traditional styles tended to be artificially resurrected. The subject matter was often selected for negative purposes—to avoid offense, to seem impressive, to fill gaps. The unsuitability of this sort of task for the artist with a romantic sense of independence is obvious, and the situation did once arise, in the case of David d'Angers, of a sculptor's choosing his own program for one of the great public buildings in Paris (the Panthéon) against the wishes of his patrons. This same sense of independence also made for difficult relations between sculptors and architects. The quarrels between the architect of the Paris Opéra and Jean-Baptiste Carpeaux were typical; what was atypical was the success of Carpeaux's festive high relief of nymphs in abandoned dance (completed in 1869).

Another type of public sculpture—the portrait, typically in bronze, erected in a town square or other public space—

DEVANEY STOCK PHOTOS

Figure 73: "Christ of the Andes" by Mateo Alonso, 1902. In the Uspallata Pass on the border between Argentina and Chile.

Commemorative portrait sculpture

flourished in the 19th century as it had not done since the first centuries AD. The first prominent sculptures of this sort commemorating nonroyal figures since antiquity seem to have appeared in Britain. The statues of Nelson by Sir Richard Westmacott erected in Liverpool and Birmingham soon after the subject's death were followed by statues of political heroes such as Fox and Pitt. By the end of the century, even relatively minor generals, philanthropists, or entrepreneurs were commemorated in this manner—almost invariably at the expense of public subscribers. The rest of Europe eventually followed this English example.

The young countries of the New World—the United States and later the republics of Latin America—commemorated with statues heroes whom they perceived as national saviours and founders. It may be that statues of Nelson excited as much patriotic sentiment as those of Washington or Bolívar, but Nelson could not embody the nation as the others did, nor certainly could any statue of a European monarch. For Europe national pride could best be promoted by an appeal to the past. Among the most remarkable public sculpture of the 19th century must certainly be counted Carlo Marochetti's "Duke Emmanuel Philibert" (1833, Turin) and Christian Daniel Rauch's "Frederick the Great" (1836–51, East Berlin) and the several statues of Joan of Arc in France. These were works of not simply historical but also topical and political significance, as indeed was the colossal "Christ of the Andes" by Mateo Alonso erected in 1902 on the border of Chile and Argentina. Abstractions were also endowed with a more urgent ideological content than in former centuries. In France, at least in the great "Triumph of the Republic" by Jules Dalou (unveiled in 1899 in the Place de la Nation), these could be animated with genuine passions. This is not true of the Statue of Liberty in New York City, which has nonetheless made an impact on the popular imagination.

Funeral sculpture

In the 19th century, funeral sculpture was as completely revolutionized as public sculpture. Whereas previously it had only really been in England that a large section of the wealthier classes had enjoyed the privilege of erecting substantial sculptured memorials, the opening up of large landscaped municipal cemeteries made this possible elsewhere. These cemeteries, of which the finest examples are in Paris and in Italy, were free from ecclesiastical censorship, and new themes quickly developed that were appropriate for an age of doubt and of desperate faith. The sentimentality and sensationalism of the annual exhibition were found here also, and so too was much exhibitionist virtuosity devoted to depicting the veiled faces and figures of ascending souls and their androgynous angelic escorts, as well as to recording bourgeois haberdashery.

This virtuosity is largely associated with Italian sculpture; and in a sense the Italians continued to dominate sculpture throughout the Western world after the death of Canova, by supplying the skilled carvers who were everywhere employed to translate into marble ideas worked out in clay. The greatest sculptors of the 19th century tended to play a smaller part than any of their predecessors in the actual carving, and the most vital sculpture of the period is preeminently plastic: when one thinks of the broken surfaces of the portrait busts by Carpeaux, for example, or of the precarious balances, open forms, and eloquent contours of Gilbert's statuettes, one thinks of wax and clay.

(N.B.P.)

Modern sculpture

19TH-CENTURY BEGINNINGS

The origins of modern art are usually traced to the mid-19th-century rejection of Academic tradition in subject matter and style by certain artists and critics. Painters of the Impressionist school that emerged in France in the late 1860s sought to free painting from the tyranny of the subject and to explore the intrinsic qualities of colour, brushwork, and form. This expansive notion of visual rendering had revolutionary effects on sculpture as well. The French sculptor Auguste Rodin found in it a new basis for life modelling and thus restored to the art

Auguste Rodin

Figure 74: "Conversation in a Garden," wax-covered plaster sculpture by Medardo Rosso, 1893. In the Galleria Nazionale d'Arte Moderna, Rome. Height 43.2 cm.

By courtesy of the Galleria Nazionale d'Arte Moderna, Rome

a stylistic integrity that it had hardly possessed for more than two centuries.

Rodin's highly naturalistic early work, "The Age of Bronze" (1877), is effective because the banal studio pose of a man leaning on a staff produced an unconventional and expressive gesture when the staff was removed. From Honoré Daumier, Rodin had learned the bold modelling of surfaces that are emotive rather than literal; the statue is only a rough approximation that avoids the definitive finish of earlier sculpture and remains in a state of becoming. Eventually, Rodin even worked with mere fragments such as broken torsos, and he enormously enlarged the range of figure composition. The mass, until then the principal vehicle of sculptural composition, was explosively opened by these methods; in contrast to earlier sculpture, which depended on the interplay of solid and void, Rodin's works are fused with the surrounding space. These methods evolved in his many works, such as "Adam" (1880), "Eve" (1881), and others, originally conceived as a part of the masterpiece of modern sculpture, "The Gates of Hell," undertaken by Rodin in 1880 and never really completed. It was inevitable that the translucent nature of the marble surface should engage the attention of Rodin, and even though he always prepared the models in clay and left the execution in stone to assistants, such marbles as "The Kiss" (1885), when properly exhibited with light partly from the rear, appear to glow with the incandescence of their passionate intensity.

(J.Hud./J.Hm.)

Although the art of Rodin appears conservative in comparison to the painting of the time, in that he continued to use literary themes while painting did not, the new style that he evolved did much to revive sculpture's significance as an expressive medium, and his importance to 20th-century sculpture can hardly be overestimated. His fresh search and revelation of the basic movements of modern life had a profound influence on the generation of European sculptors who followed him.

Among Rodin's contemporaries, Edgar Degas, whose sculpture, begun in the 1880s, was an intimate study of movement and light, in several respects predicts 20th-century developments. Rodin's Italian counterpart, Medardo Rosso, lived in Paris during the 1880s; his work was known and owned by Rodin (Figure 74). Less gifted than Rodin but interested in the same problems, Rosso used wax in such a way that light was suffused through sensitively modelled portraits, and labile forms were created to express the flux that he felt was a condition of modern life. In Italy Rosso influenced Arturo Martini and through him Giacomo Manzù, Marino Marini, and Alberto Viani.

THE 20TH CENTURY

Rodin's successors

The ablest of Rodin's many pupils were Émile-Antoine Bourdelle and Charles Despiau. Bourdelle's "Héraklès Archer" (1910) is an attempt to continue Rodin's active postures; but the results are melodramatic, and the forms are heavy and less sensitively modelled. Despiau, who

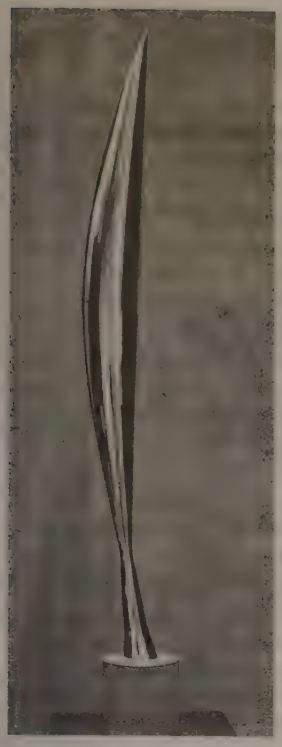


Figure 75: "Bird in Space," polished bronze sculpture by Constantin Brancusi, 1928? In the Museum of Modern Art, New York City. Height 1.37 m.

Collection, The Museum of Modern Art, New York, given anonymously

was director of Rodin's shop from 1907 to 1914, also responded to the interest in Classicism; his best work, "Girl from the Landes" (1904), was a balance of individual traits in the Rodin tradition, combined with graceful poses and well-rounded forms.

Two of the many other young sculptors attracted to Paris by Rodin's fame were Wilhelm Lehmbruck and Constantin Brancusi. Lehmbruck's early work has the soft modelling by touches of clay characteristic of the time, as in his "Mother and Child" (1907) and "Bust of a Woman" (1910). Brancusi's "Sleeping Muse" (1908) and the small "Bust of a Boy with Head Inclined" (1907) reflect Rodin's later interests in the expressiveness of modelling as opposed to strenuous gesture. Pablo Picasso and Henri Matisse were also early disciples of Rodin, as was Jacob Epstein, particularly in his naturalistic and psychologically incisive portraits.

Collection, The Museum of Modern Art, New York, Lillie P. Bliss Bequest

Figure 76: "Unique Forms of Continuity in Space," bronze sculpture by Umberto Boccioni, 1913. In the Museum of Modern Art, New York City. Height 1.1 m.

Avant-garde sculpture (1909–20). In the second decade of the 20th century the tradition of body rendering extending from the Renaissance to Rodin was shattered, and the Cubists, Brancusi, and the Constructivists emerged as the most influential forces. Cubism, with its compositions of imagined rather than observed forms and relationships, had a similarly marked influence.

One of the first examples of the revolutionary sculpture is Picasso's "Woman's Head" (1909). The sculptor no longer relied upon traditional methods of sculpture or upon his sensory experience of the body; what was given to his outward senses of sight and touch was dominated by strong conceptualizing. The changed and forceful appearance of the head derives from the use of angular planar volumes joined in a new syntax independent of anatomy. In contrast to traditional portraiture, the eyes and mouth are less expressive than the forehead, cheeks, nose, and hair. Matisse's head of "Jeanette" (1910–11) also partakes of a personal reproportioning that gives a new vitality to the less mobile areas of the face. Likewise influenced by the Cubists' manipulation of their subject matter, Alexander Archipenko in his "Woman Combing Her Hair" (1915) rendered the body by means of concavities rather than convexities and replaced the solid head by its silhouette within which there is only space.

Brancusi also abandoned Rodin's rhetoric and reduced the body to its mystical inner core. His "Kiss" (1908), with its two blocklike figures joined in symbolic embrace, has a concentration of expression comparable to that of primitive art but lacking its spiritualistic power. In this and subsequent works Brancusi favoured hard materials and surfaces as well as self-enclosed volumes that often impart an introverted character to his subjects. His bronze "Bird in Space" became a *cause célèbre* in the 1920s when U.S. customs refused to admit it duty free as a work of art (Figure 75).

Brancusi's "Bird in Space"

Raymond Duchamp-Villon began as a follower of Rodin, but his portrait head "Baudelaire" (1911) contrasts with that by his predecessor in its more radical departure from the flesh; the somewhat squared-off head is molded by clear, hard volumes. His famous "Horse" (1914), a coiled, vaguely mechanical form bearing little resemblance to the animal itself, suggests metaphorically the horsepower of locomotive drive shafts and, by extension, the mechanization of modern life. Duchamp-Villon may have been influenced by Umberto Boccioni, one of the major figures in the Italian Futurist movement and a sculptor who epitomized the Futurist love of force and energy deriving from the machine. In "Unique Forms of Continuity in Space" (Figure 76) and "Head + House + Light" (1911), he carried out his theories that the sculptor should model objects as they interact with their environment, thus revealing the dynamic essence of reality.

Jacques Lipchitz came to Cubism later than Archipenko and Duchamp-Villon, but after mastering its meaning he produced superior sculpture. In 1913, after several years of conservative training, he made a number of small bronzes experimenting with the compass curve and angular planes. They reveal an understanding of the Cubist reconstitution of the bodies in an impersonal quasi-geometric armature over which the artist exercised complete autonomy. Continuing to work in this fashion, he produced "Man with a Guitar" (Figure 77), and "Standing Figure" (1915), in which voids are introduced, while in the early 1920s he developed freer forms more consistently based on curves.

Lehmbruck's mature style emerged in the "Kneeling Woman" (1911) and "Standing Youth" (1913), in which his gothicized, elongated bodies with their angular posturings and appearance of growing from the earth give

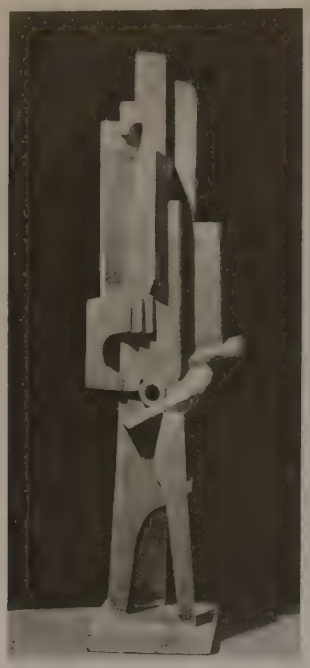


Figure 77: "Man with a Guitar," limestone sculpture by Jacques Lipchitz, 1916. In the Museum of Modern Art, New York City. Height 97.2 cm.

Collection, The Museum of Modern Art, New York, Mrs. Simon Guggenheim Fund (by exchange)

expression to his notions of modern heroism. In contrast to this spiritualized view is his "The Fallen" (1915–16), intended as a compassionate memorial for friends lost in the war.

Constructivism and Dada. Between 1912 and 1914 there emerged an antisculptural movement, called Constructivism, that attacked the false seriousness and hollow moral ideals of academic art. The movement began with the relief fabrications of Vladimir Tatlin in 1913. The Constructivists and their sympathizers preferred industrially manufactured materials, such as plastics, glass, iron, and steel, to marble and bronze. Their sculptures were not formed by carving, modelling, and casting but by twisting, cutting, welding, or literally constructing: thus the name Constructivism.

Unlike traditional figural representation, the Constructivists' sculpture denied mass as a plastic element and volume as an expression of space; for these principles they substituted geometry and mechanics. In the machine, where the Futurists saw violence, the Constructivists saw beauty. Like their sculptures, it was something invented; it could be elegant, light, or complex, and it demanded the ultimate in precision and calculation.

Sculpture as invention

Seeking to express pure reality, with the veneer of accidental appearance stripped away, the Constructivists fabricated objects totally devoid of sentiment or literary association; Naum Gabo's work frequently resembled mathematical models, and several Constructivist sculptures, such as those by Kazimir Malevich and Georges Vantongerloo, have the appearance of architectural models. The Constructivists created, in effect, sculptural metaphors for the new world of science, industry, and production; their aesthetic principles are reflected in much of the furniture, architecture, and typography of the Bauhaus.

A second important offshoot of the Cubist collage was the fantastic object or Dadaist assemblage. The Dadaist movement, while sharing Constructivism's iconoclastic vigour, opposed its insistence upon rationality. Dadaist assemblages were, as the name suggests, "assembled" from materials lying about in the studio, such as wood, cardboard, nails, wire, and paper; examples are Kurt Schwitters' "Rubbish Construction" (1921) and Marcel Duchamp's "Disturbed Balance" (1918). This art generally exalted the accidental, the spontaneous, and the impulsive, giving free play to associations. Its paroxysmal and negativist tenor led its subscribers into other directions, but Dadaism formed the basis of the imaginative sculpture that emerged in the later 1920s.

The Dadaist assemblage

Conservative reaction (1920s). In the 1920s modern art underwent a reaction comparable to the changes experienced by society as a whole. In the postwar search for security, permanence, and order, the earlier insurgent art seemed to many to be antithetical to these ends, and certain avant-garde artists radically changed their art and thought. Lipchitz' portraits of "Gertrude Stein" (1920) and "Berthe Lipchitz" (1922) return volume and features to the head but not an intimacy of contact with the viewer. Tatlin and Alexander Rodchenko broke with the Constructivists around 1920. Jacob Epstein developed some of his finest naturalistic portraiture in this decade. Rudolph Belling abandoned the mechanization that had characterized his "Head" (1925) in favour of musculature and individual identity in his statue of "Max Schmeling" of 1929. Matisse's reclining nudes and the "Back" series of 1929 show less violently worked surfaces and more massive and obvious structuring.

Aristide Maillol continued refining his relaxed and uncomplicated female forms with their untroubled, stolid surfaces. In Germany, Georg Kolbe's "Standing Man and Woman" of 1931 seems a prelude to the Nazi health cult, and the serene but vacuous figures of Arno Breker, Karl Albiker, and Ernesto de Fiori were simply variations on a studio theme in praise of youth and body culture. In the United States adherents of the countermovement included William Zorach, Chaim Gross, Adolph Block, Paul Manship, and Wheeler Williams.

Sculpture of fantasy (1920–45). One trend of Surrealist or Fantasist sculpture of the late 1920s and the 1930s consisted of compositions made up of found objects, such as Meret Oppenheim's "Object, Fur Covered Cup" (1936). As with Dadaist fabrications, the unfamiliar conjunction of familiar objects in these assemblies was dictated by impulse and irrationality and could be summarized by Isidore Ducasse's often-quoted statement, "Beautiful . . . as the chance meeting on a dissecting table of a sewing machine with an umbrella."

Of greater artistic importance was the sculpture of a second group that included Alberto Giacometti, Jean Arp, Lipchitz, Henry Moore, Barbara Hepworth, Picasso, Julio González, and Alexander Calder. Although these sculptors were sometimes in sympathy with Surrealist objectives, their aesthetic and intellectual concerns prohibited a more consistent attachment. Their art, derived from visions, hallucinations, reverie, and memory, might best be called the sculpture of fantasy. Giacometti's "Palace at 4 A.M." (Figure 78), for example, interprets the artist's vision not in terms of the external public world but in an enigmatic,

By courtesy of the Museum of Modern Art, New York

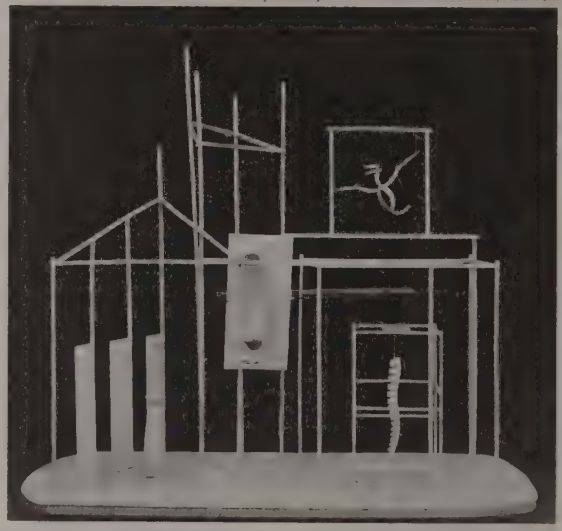


Figure 78: "The Palace at 4 A.M.," mixed media (wood, glass, wire, and string) by Alberto Giacometti, 1932–33. In the Museum of Modern Art, New York City. 63.5 × 71.8 × 40 cm.

private language. Moore's series of "Forms" suggest shapes in the process of forming under the influence of each other and the medium of space. The appeal of primitive and ancient ritual art to Moore, the element of surprise in children's toys for Calder, and the wellsprings of irrationality from which Arp and Giacometti drank were for these men the means by which wonder and the marvelous could be restored to sculpture. While their works are often violent transmutations of life, their objectives were peaceful, "... to inject into the vain and bestial world and its retinue, the machines, something peaceful and vegetative." ([Jean] Hans Arp, *On My Way,* Documents of Modern Art, vol. 6, p. 123, George Wittenborn, Inc., New York, 1948.)

Other sculpture (1920–45). The sculpture of Moore, Gaston Lachaise, and Henri Laurens during the 1920s and '30s included mature, ripe human bodies, erogenic images reminiscent of Hindu sculpture, appearing inflated with breath rather than supported by skeletal armatures. Lachaise's "Montagne" (1934–35) and Moore's reclining nudes of the '30s are identifications with earth, growth, vital rhythm, and silent power. Prior to Moore and the work of Archipenko, Boccioni, and Lipchitz, space had been a negative element in figure sculpture; in Moore's string sculptures and Lipchitz' transparencies of the 1920s, it became a prime element of design.

Moore's reclining nudes

Lipchitz' figure style of the late 1920s and '30s is inseparable from his emerging optimistic humanism. His concern with subject matter began with the ecstatic "Joy of Life" (1927). Thereafter his seminal themes were of love and security and assertive passionate acts that throw off the inertia of his Cubist figures. In the "Return of the Prodigal Son" (1931), for example, strong, facetted curvilinear volumes weave a pattern of emotional and aesthetic accord between parent and child.

The American sculptor John B. Flannagan rendered animal forms as well as the human figure in a simple, almost naive style. His interest in what he called the "profound subterranean urges of the human spirit in the whole dynamic life process, birth, growth, decay and death" (quoted in Carl Zigrosser, *Catalog for the Exhibition of the Sculpture of John B. Flannagan,* p. 8, The Museum of Modern Art, New York, 1942) resulted in "Head of a Child" (1935), "New One" (1935), "Not Yet" (1940), and "The Triumph of the Egg" (1941).

Somewhat more mystical are Brancusi's "Beginning of the World" (1924), "Fish" (1928–30), and "The Seal" (1936). As with Flannagan, the recurrent egg form in Brancusi's art symbolizes the mystery of life. Nature in motion is the subject of Alexander Calder's mobiles, such as "Lobster Trap and Fish Tail" (1939) and others suggesting the movement of leaves, trees, and snow. In the history of sculpture there is no more direct or poetic expression of nature's rhythm.

Developments after World War II. "The modern artist is the counterpart in our time of the alchemist-philosopher who once toiled over furnaces, alembics and crucibles, ostensibly to make gold, but who consciously entered the most profound levels of being, philosophizing over the melting and mixing of various ingredients" (Ibram Lassaw, quoted in Lawrence Campbell in *Art News,* p. 66, The Art Foundation Press, New York, March 1954). While work in the older mediums persisted, it was the welding, soldering, and cutting of metal that emerged after 1945 as an increasingly popular medium for sculpture. The technical and expressive potential of uncast metal sculpture was carried far beyond the earlier work of González and Picasso.

The appeal of metal is manifold. It is plentifully available from commercial supply houses; it is flexible and permanent; it allows the artist to work quickly; and it is relatively cheap compared to casting. Industrial metals also relate modern sculpture physically, aesthetically, and emotionally to its context in modern civilization. As the American sculptor David Smith has commented, "Possibly steel is so beautiful because of all the movement associated with it, its strength and functions. Yet it is also brutal, the rapist, the murderer and death-dealing giants are also its offspring" (quoted in Garola Giedion-Welcker, *Contemporary Sculpture,* Documents of Modern Art, vol. 12, p. 123, George Wittenborn, Inc., New York, 1955).

The technique of metal sculpture

The basic tool of the metal sculptor is the oxyacetylene torch, which achieves a maximum temperature of 6,500° F (3,600° C; the melting point of bronze is 2,000° F). The intensity and size of the flame can be varied by alternating torch tips. In the hands of a skilled artist the torch can cut or weld, harden or soften, colour and lighten or darken metal. Files, hammers, chisels, and jigs are also used in shaping the metal, worked either hot or cold. The sculptor may first construct a metal armature that he then proceeds to conceal or expose. He builds up his form with various metals and alloys, fusing or brazing them, and may expose parts or the whole to the chemical action of acids. This type of work requires constant control, and many sculptors work out and guard their own recipes.

Other sculptors such as Peter Agostini, George Spaventa, Peter Grippe, David Slivka, and Lipchitz, who were interested in bringing spontaneity, accident, and automatism into play, returned to the more labile media of wax and clay, with occasional cire-perdue casting, which permit a very direct projection of the artist's feelings. By the nature of the processes such work is usually on a small scale.

A number of artists brought new technique and content to the Dadaist form of the assemblage. Among the most important was the American Joseph Cornell, who combined printed matter and three-dimensional objects in his intimately sealed, often enigmatic "boxes."

Another modern phenomenon, seen particularly in Italy, France, and the United States, was the revival of relief sculpture and the execution of such works on a large scale, intended to stand alone rather than in conjunction with a building. Louise Nevelson, for example, typically employed boxes as container compartments in which she carefully disposed an assortment of forms and then painted them a uniform colour. In Europe the outstanding metal reliefs were those by Alberto Burri, Gio and Arnaldo Pomodoro, César (Figure 79), Zoltán Kemény, and Manuel Rivera.

New views of nature. Development of metal sculpture, particularly in the United States, led to fresh interpretations of the natural world. In the art of Richard Lippold and Ibram Lassaw, the search for essential structures took the form of qualitative analogies. Lippold's "Full Moon" (1949–50) and "Sun" (1953–56; commissioned by the Metropolitan Museum of Art, New York City, to hang in its room of Persian carpets) show an intuition of a basic regularity, precise order, and completeness that underlies the universe. Lassaw's comparable interest in astronomical phenomena inspired his "Planets" (1952) and "The Clouds of Magellan" (1953).

In contrast to the macrocosmic concern of these two artists were the interests of sculptors such as Raymond Jacobson, whose "Structure" (1955) derived from his study of honeycombs. Using three basic sizes, Jacobson constructed his sculpture of hollowed cubes emulating the modular, generally regular but slightly unpredictable formal quality of the honeycomb.

Isamu Noguchi's "Night Land" is one of the first pure landscapes in sculpture. David Smith's "Hudson River

Figure 79: "The Man of Saint-Denis," by César (Baldaccini), 1958. In the Tate Gallery, London. Height 50.8 cm.

Landscape" (1951), Theodore J. Roszak's "Recollections of the Southwest" (1948), Louise Bourgeois's "Night Garden" (1953), and Leo Amino's "Jungle" (1950) are later examples.

In the 1960s a number of sculptors, particularly in the United States, began to experiment with using the natural world as a kind of medium rather than a subject. Among the more notable examples were the American Robert Smithson, who frequently employed earth-moving equipment to alter natural sites, and the Bulgarian-born Christo, whose "wrappings" of both natural and man-made structures in synthetic cloth generated considerable controversy. The name environmental sculpture has come to denote such works, together with other sculptures that constitute self-contained environments.

The human figure. Since figural sculpture moved away from straightforward imitation, the human form has been subjected to an enormous variety of interpretations. The thin, vertical, Etruscan idol-like figures developed by Giacometti showed his repugnance toward rounded and smooth body surfaces or strong references to the flesh. His men and women do not exist in felicitous concert with others; each form is a secret sanctum, a maximum of being wrested from a minimum of material. Reg Butler's work (*e.g.*, "Woman Resting" [1951]) and that of David Hare ("Figure in a Window" [1955]) treat the body in terms of skeletal outlines. Butler's figures partake of nonhuman qualities and embody fantasies of an unsentimental and aggressive character; the difficulties and tensions of existence are measured out in taut wire armatures and constricting malleable bronze surfaces. Kenneth Armitage and Lynn Chadwick, two other British sculptors, make the clothing a direct extension of the figure, part of a total gesture. In his "Family Going for a Walk" (1953), for example, Armitage creates a fanciful screenlike figure recalling wind-whipped clothing on a wash line. Both Chadwick and Armitage transfer the burden of expression from human limbs and faces to the broad planes of the bulk of the sculpture. Chadwick's sculptures are often illusive hybrids suggesting alternately impotent De Chirico-like figures or animated geological forms.

Luciano Minguzzi admired the amply proportioned feminine form. Minguzzi's women (*e.g.*, "Woman Jumping Rope" [1954]) may exert themselves with a kind of playful abandon. Marini's women (*e.g.*, "Dancer" [1949]) enjoy a stately passivity, their quiescent postures permitting a contrapuntal focus on the graceful transition from the slender extremities to the large, compact, voluminous torso, with small, rich surface textures.

The segmented torso, popular with Arp, Laurens, and Picasso earlier, continued to be reinterpreted by Alberto Viani, Bernard Heiliger, Karl Hartung, and Raoul Hague. The emphasis of these sculptors was upon more subtle, sensuous joinings that created self-enclosing surfaces.

A.C. Cooper Ltd

Figure 80: "Family Going for a Walk," bronze sculpture by Kenneth Armitage, 1953. In a private collection. Height 74 cm.

Viani's work, for example, does not glorify body culture or suggest macrocosmic affinities as does an ideally proportioned Phidian figure; his torsos are seen in a private way, as in his "Nude" (1951), with its large body and golf ball-sized breasts.

Among the most impressive figure sculptures made in the United States in the late 1950s were those by Seymour Lipton. Their large-scale, taut design and provocative interweaving of closed and open shapes restore qualities of mystery and the heroic to the human form.

The American George Segal emerged from the Pop movement of the 1950s and '60s as a major figurative sculptor. His plaster casts from live models, usually left white and indistinctly featured, are often situated in mundane settings of actual furniture or other objects.

The works of the French-born American artist Marisol contrast sharply with Segal's in their boxlike forms, onto which highly individualized features are usually painted. In the 1970s and '80s, Duane Hanson, another American, took Segal's live-model casting technique a step further with his startlingly naturalistic, fully pigmented cast fibreglass figures.

Archaizing, idol making, and religious sculpture. After World War II several sculptors became interested in the art of early Mediterranean civilizations. The result was a conscious archaizing of the human form with the intent of recapturing qualities of Cycladic idols, early Greek and Egyptian statuary, and some aspects of late Roman art.

Moore's admiration for archaic Greek sculpture produced "Draped Reclining Figure" (1952), which shows his return to the solid form and the suggestion of power and force by using drapery as a tense foil for the volumes that press against it. His "King and Queen" (1952–53) resulted from further excursions into the archaic Greek myth world.

The interest in recreating idols or totems was continued by Arp in his "Idol" (1950) and by Noguchi in his Stone Age-type sculptures for the Connecticut General Life Insurance Company (Hartford). By creating presences that elude rational definition, these artists restored to art its ancient aura of myth, mystery, and magic in an age that consistently disclaims their existence.

Modern religious sculpture

The argument that modern sculpture is inappropriate for religious requirements is disproved by works of Lipchitz, Lassaw, and Herbert Ferber. In keeping with the Jewish preference for nonfigural art, Ferber's ". . . and the bush was not consumed" (1951), commissioned by a synagogue in Millburn, New Jersey, comprises clusters of branches and boldly shaped weaving flames, invisibly suspended in a powerful and intimate vision that absorbs its viewers with its hypnotic rhythm. Lassaw's "Pillar of Fire," for the exterior of a synagogue in Springfield, Massachusetts, also has a mesmerizing pattern recalling the illusory images sometimes seen in flames. Lipchitz' sculpture of the "Virgin of Assy" (1948–54) was commissioned for the Catholic church at Assy, France.

Moreover, an increasing number of gifted sculptors are providing handsome liturgical objects and decorations, such as Harry Bertoia's shimmering reredos, Lipton's work for a synagogue in Tulsa, Oklahoma, and Roszak's sculptured spire for Kresge Chapel on the campus of the Massachusetts Institute of Technology, Cambridge.

Public and private memorials. After World War II there was a flood of public memorial sculpture, and in Europe especially many of the commissions were carried out by modern sculptors. A striking war memorial in Italy is Mirko Basaldella's gate for the monument to the Roman hostages killed in the Ardeatine Caves (1951). For its full effect the gate must be seen in connection with the rugged masonry wall to which it is attached. The gate was cast in metal and fashioned in a tangled, thicket-like pattern that suggests the painfully difficult passage from life to death for those who died in the caves.

Another imposing memorial is Ossip Zadkine's monument to the bombing of Rotterdam, a figure recoiling from the violence that descended from the sky. In Moore's "Warrior with a Shield" a soldier defiantly raises his shield and mutilated body toward the ill-starred heavens during the Battle of Britain. Epstein's public monument to "Social Consciousness" (1952–53), in Fairmount Park,

Figure 81: "Social Consciousness," bronze sculpture by Jacob Epstein, 1952–53. In Fairmount Park, Philadelphia. Height 3.9 m.

By courtesy of Philadelphia Museum of Art, Ellen Phillips Samuel Memorial, Fairmount Park Art Association

Philadelphia, treats the helplessness of those confronted with pressures over which they have no control. In contrast to the invulnerable champions of academic art, these sculptures image the hero in distress.

Modern sculpture and architecture

Other developments. Despite the rapid and exciting developments in both architecture and sculpture, the two have seldom been meaningfully and integrally united. The architecture of Le Corbusier, Frank Lloyd Wright, Pier Luigi Nervi, Ludwig Mies van der Rohe, and others occasionally shows strong sculptural qualities, but relatively rarely were their surfaces planned to receive sculpture. Freestanding sculptures such as those created by Gabo, Pevsner, De Rivera, Calder, and Noguchi have been used to provide intimacy and visual relief from the severity of the "cult of the cube" in architecture. The architectural firm of Skidmore, Owings & Merrill successfully used Bertoia's brilliant screens and Noguchi's sculptures and garden ideas; Roszak's "Eagle" for the American embassy in London and Moore's changeable reliefs on the London Time and Life Building held out hope for further thoughtful integration of the arts.

Also of great moment is the phenomenon of the sculptor-designer who has produced important changes in furniture and industrial design. Max Bill's school in Ulm, Germany, showed great promise. Playground facilities have been revolutionized by such designs as those made by Noguchi for Creative Playthings Inc. in the United States and the slides, hollowed forms, and climb apparatus of Egon Moeller-Nielson for parks in Stockholm. Noguchi, Moholy-Nagy, Bill, Bertoia, and many other modern artists contributed to the breakdown of the distinction between the object of utility and the work of art. Not since Gothic times has sculpture shown such promise of becoming an extensive and important part of human existence. (A.E.El./Ed.)

In Italy, traditional trends in sculpture are reflected in the brilliant accomplished modelling of Giacomo Manzù; Marino Marini, devoting himself almost entirely to the single theme of horse and rider, gave a bald realistic style an oddly apocalyptic force. The rough-hewn monumentality of the figures of the Austrian carver Fritz Wotruba is characteristic of this phase. Joannis Avramidis, also working in Vienna, turned figures into clusters of simplified formal echoes; the third sculptor of the Viennese group, Rudolf Hoflehner, who worked in iron, transformed them into symbolic presences. The segmental iron sculpture of the Spaniard Eduardo Chillida deals with a more limited and powerful range of forms.

Robert Rauschenberg in the United States sought to place his subtly calculated "combines" in the gap between reality and art, contrasting the significance of paint with the borrowed imagery and objects that are juxtaposed to it. Another American, Claes Oldenburg, began by reconstructing common things out of the random pictorial substance of Abstract Expressionism; his later reconstructions of the rigid furniture of life are tailored out of limp plastic sheeting, and the paradox oddly extends one's knowledge of the objective world.

In the reliefs of the Venezuelan Jesús Raphael Soto, the shifting paradoxes of vision are given a delicate order. Aside from this, the widespread work in kinetic mediums, such as that of Nicholus Takis, during the 1960s formed a separate genre, winking and shuddering on its own, most nearly linked to the Surrealist tradition.

Other sequels of the general rationalization and concentration of artistic means have been more fertile. In the hands of the U.S. painters Kenneth Noland and Frank Stella, painting discovered new shapes, both within the rectangular canvas and beyond it. The new value that was given to the painted plane did not benefit painting only. The British painter Richard Smith deployed it in three dimensions in painted constructions that re-create impressions of commercial packaging in terms of the spatial imagination of the arts. Sculpture, reequipped with colour, developed remarkably, and Anthony Caro led a group of British sculptors in exploration of spatial modulation and formal analogy. (La.G.)

BIBLIOGRAPHY

General: An excellent general history of world art is HUGH HONOUR and JOHN FLEMING, *A World History of Art* (1982; U.S. title, *The Visual Arts: A History*), which examines sculpture in relation to the other arts. H.W. JANSON, *History of Art* (1962; 2nd ed., 1977), is also recommended. Among books that discuss sculpture of many periods, RUTH BUTLER, *Western Sculpture: Definitions of Man* (1975), is unusually valuable. So, too, is F. DAVID MARTIN, *Sculpture and Enlivened Space* (1981). For the techniques of sculpture see W. VERHELST, *Sculpture: Tools, Materials, and Techniques* (1973); and RUDOLF WITTKOWER, *Sculpture* (1977). The making of bronze sculptures, omitted from the latter, is brilliantly elucidated by JENNIFER MONTAGU, *Bronzes* (1963, reissued 1972). ERWIN PANOFSKY, *Tomb Sculpture* (1964), traces from ancient Egypt to about 1800 some of the major themes of one very important class of Western sculpture.

Ancient Mediterranean: Sculpture in the early civilizations of southern Europe is seldom studied separately, but it is featured in the following general works: JOHN BOARDMAN, *Pre-Classical* (1967, reissued 1979); R.W. HUTCHINSON, *Prehistoric Crete* (1962); A. ARRIBAS, *The Iberians* (1964); N.K. SANDARS, *Prehistoric Art in Europe* (1968); and SPYRIDON MARINATOS, *Crete and Mycenae* (1960).

Greek, Hellenistic, Etruscan, and Roman art: An authoritative and comprehensive account of ancient Greek art (which, for the most part, means Greek sculpture) is MARTIN ROBERTSON, *A History of Greek Art* (1975). For a succinct introduction to sculpture only, see JOHN BARRON, *Introduction to Greek Sculpture* (1981, reissued 1984). For the Archaic period, G.M.A. RICHTER, *Archaic Greek Art Against Its Historical Background* (1949), is still valuable; for the so-called Classical period, BRUNILDE S. RIDGWAY, *Fifth Century Styles in Greek Sculpture* (1981), is a good detailed guide; and for the later periods, MARGARETE BIEBER, *The Sculpture of the Hellenistic Age*, 2nd rev. ed. (1981), is highly useful. For the ancient literature on art, see J.J. POLLITT, *The Art of Greece 1400–31 B.C.: Sources and Documents* (1965). Etruscan sculpture is best discussed in OTTO J. BRENDEL, *Etruscan Art* (1978). Sculpture features prominently in the most lively general books on Roman art: R. BIANCHI BANDINELLI, *Rome: The Centre of Power* (1970; originally published in Italian, 1969), and *Rome: The Late Empire* (1971); and RICHARD BRILLIANT, *Roman Art* (1974). Of more limited scope but great interest is JOCELYN M.C. TOYNBEE, *Art in Roman Britain* (1962). See also J.J. POLLITT, *The Art of Rome c. 753 BC–AD 337: Sources and Documents* (1966, reissued 1983).

Early Christian and early medieval: Good general surveys of the early Christian period that include some discussion of sculpture are ERNST KITZINGER, *Byzantine Art in the Making* (1977); JOHN BECKWITH, *The Art of Constantinople*, 2nd ed. (1968); ANDRÉ GRABAR, *The Beginnings of Christian Art: 200–395* (1967, originally published in French, 1966); STEVEN RUNCIMAN, *Byzantine Style and Civilization* (1975); and CYRIL A. MANGO, *The Art of the Byzantine Empire 312–1453: Sources and Documents* (1972). This last volume, together with ERNST KITZINGER, *Early Medieval Art* (1940; rev. ed., 1983), concerns also the early medieval period. Among more specialized studies

of sculpture in the early Christian period, JOHN BECKWITH, *Coptic Sculpture* (1963); and JOSEPH NATANSON, *Early Christian Ivories* (1953), should be mentioned. For general information on the early medieval period, see PETER LASKO, *Ars Sacra 800–1200* (1972); GEORGE HENDERSON, *Early Medieval* (1972); and GEORGE ZARNECKI, *Art of the Medieval World* (1975). Valuable studies specifically on sculpture include GEORGE H. CRICHTON, *Romanesque Sculpture in Italy* (1954); HERMANN LEISINGER, *Romanesque Bronzes* (1956); FRITZ SAXL, *English Sculptures of the Twelfth Century* (1954); and M.F. HEARN, *Romanesque Sculpture: The Revival of Monumental Stone Sculpture in the Eleventh and Twelfth Centuries* (1981).

Gothic: Many of the ideas expressed in this section of the article are treated at greater length in ANDREW MARTINDALE, *Gothic Art* (1967). General studies of Gothic art include GEORGE HENDERSON, *Gothic* (1967); JOAN EVANS (ed.), *The Flowering of the Middle Ages* (1966, reissued 1984); and JOHAN HUIZINGA, *The Waning of the Middle Ages* (1924, reissued 1976; 12th Dutch ed., 1973). For the imagery of the period, the reader is referred to ÉMILE MÂLE, *The Gothic Image: Religious Art in France of the Thirteenth Century* (1958, reissued 1972; trans. of 3rd French ed., 1910), and *Religious Art from the Twelfth to the Eighteenth Century* (1949, reissued 1970; originally published in French, 1945). A useful anthology of the literary sources of the period is TERESA G. FRISCH, *Gothic Art 1140–1450* (1971). For a general treatment of English Gothic sculpture, see LAWRENCE STONE, *Sculpture in Britain: The Middle Ages,* 2nd ed. (1972); for France, MARCEL AUBERT, *La Sculpture française au moyen âge* (1947); and for Italy, JOHN POPE-HENNESSY, *Italian Gothic Sculpture,* 2nd ed. (1972).

Renaissance: There are numerous general books on Renaissance art, especially on Renaissance art in Italy, but sculpture is seldom adequately discussed in them. The best introduction to the sculpture is JOHN POPE-HENNESSY, *Italian Renaissance Sculptures,* 2nd ed. (1971). As a succinct guide to the sculpture in Florence, the most consistently important centre in Europe at this time, CHARLES AVERY, *Florentine Renaissance Sculpture* (1970), is recommended. Renaissance sculpture in northern Europe is discussed in ANTHONY BLUNT, *Art and Architecture in France: 1500–1700* (1953); WOLFGANG STECHOW, *Northern Renaissance Art: 1400–1600* (1966); GERT VON DER OSTEN and HORST VEY, *Painting and Sculpture in Germany and the Netherlands: 1500–1600* (1969); and MICHAEL BAXANDALL, *The Limewood Sculptors of Renaissance Germany* (1980). For Spain and Portugal, see GEORGE KUBLER and MARTIN S. SORIA, *Art and Architecture in Spain and Portugal and Their American Dominions: 1500–1800* (1959).

Baroque and Rococo: The best brief general discussion of Western art of this period is MICHAEL KITSON, *The Age of Baroque* (1966, reissued 1976), which includes some consideration of sculpture. For Italian Baroque sculpture, a better guide than POPE-HENNESSY (above) is provided by the sections on sculpture in RUDOLF WITTKOWER, *Art and Architecture in Italy: 1600–1750,* 3rd rev. ed. (1973, reissued 1982). ROBERT ENGGASS, *Early Eighteenth-Century Sculpture in Rome,* 2 vol. (1976); and the first two volumes (1977 and 1981) of FRANÇOIS SOUCHAL, *French Sculptors of the 17th and 18th Centuries,* must also be mentioned. For 18th-century France, the sections by Michael Levey on sculpture in MICHAEL LEVEY and WEND GRAF KALNEIN, *Art and Architecture of the Eighteenth Century in France* (1972), are excellent. For English sculpture, see the admirable account in MARGARET WHINNEY, *Sculpture in Britain: 1530–1830* (1964). For Spain, Portugal, and Latin America, see KUBLER and SORIA (above); HAROLD E. WETHEY, *Colonial Architecture and Sculpture in Peru* (1949, reprinted 1971); and PAL KELEMEN, *Baroque and Rococo in Latin America* (1951).

Neoclassicism and the 19th century: An excellent general account of Neoclassicism, which includes much of value on sculpture, is HUGH HONOUR, *Neoclassicism* (1977). For England, see DAVID G. IRWIN, *English Neoclassical Art* (1966); BENEDICT READ, *Victorian Sculpture* (1982); SUSAN BEATTIE, *The New Sculpture* (1983); and WHINNEY (above). For France and Italy, see GERARD HUBERT, *La Sculpture dans l'Italie Napoléonienne* (1964); JANE VAN NIMMEN and RUTH MIROLLI, *Nineteenth Century French Sculpture* (1971), an admirable introduction; and PETER FUSCO and H.W. JANSON (eds.), *The Romantics to Rodin* (1980), also a good introduction. A superb general introduction—perhaps the only truly comprehensive one—to Western sculpture of the 19th century is H.W. Janson's contribution to ROBERT ROSENBLUM and H.W. JANSON, *Art of the Nineteenth Century* (1984; U.S. title, *19th Century Art*).

Modern: There are numerous general introductions to modern art, but most give little space to sculpture. The best books devoted to modern sculpture are ALBERT E. ELSEN, *Modern European Sculpture: 1918–1945* (1979); HERBERT READ, *A Concise History of Modern Sculpture* (1964); and FRED LICHT, *Sculpture: 19th and 20th Centuries* (1967). Some recent developments are described in ALLEN KAPROW, *Assemblage: Environments and Happenings* (1966); and UDO KULTERMANN, *The New Sculpture* (1968; originally published in German, 1967). For a prominent sculptor's compelling but contentious account of what sculpture consists of, see WILLIAM TUCKER, *The Language of Sculpture* (1977).

(N.B.P.)

Sensory Reception

Sensory reception is the means by which an organism detects and responds to changes in its external or internal environment. Organisms have a variety of sensory structures that respond to different stimuli, such as light, pressure, or chemicals, all of which are forms of energy. Once excited, these sensory receptors convert the energy of the stimulus into a behavioral response of the organism.

In unicellular organisms environmental signals are received by specialized organelles, such as light-sensitive eyespots or hairlike cilia sensitive to mechanical disturbances. In multicellular organisms sensory signals can be transmitted from a receptor organ to other parts of the body by specialized cells. For example, in all higher animals sensory reception is the special function of sensory neurons, which convert, or transduce, a stimulus into the electrochemical activity of nerve impulses. These impulses are transmitted to the brain, where they are processed and interpreted. In general, the more highly evolved the organism, the more complex is its sensory apparatus.

This article examines the sensory capacities of living organisms, explains the mechanisms of sensory function, and considers the adaptive advantages of sensing. Human sensory reception is presented in separate sections. Diseases and impairments of seeing and hearing are also discussed.

This article is divided into the following major sections:

Animal sensory reception 115
- Nature and functions of sensory systems 115
  - Classification of sensory systems
  - Evolution of sensory systems
  - Integration of sensory information
- Mechanoreception 117
  - Reception of external mechanical stimuli
  - Reception of internal mechanical stimuli
  - Maintenance of equilibrium
- Thermoreception 124
  - General properties of thermoreceptors
  - Thermoreceptors in invertebrates
  - Thermoreceptors in vertebrates
- Chemoreception 128
  - Classes of chemoreceptors
  - Adaptive functions of chemoreception
  - Chemoreceptors in lower invertebrates
  - Arthropod chemoreceptors
  - Chemoreception in the vertebrates
  - Theories of chemoreceptor action
- Photoreception 140
  - The optical properties of eyes
  - The properties of photoreceptors
  - Physiological response of photoreceptors
- Sound reception 152
  - Organs of sound reception in invertebrates
  - Sound reception in vertebrates—auditory mechanisms of fishes and amphibians
  - Auditory structures of reptiles
  - Hearing in birds
  - Hearing in mammals

Human sensory reception 164
- General considerations of sensation 165
  - Basic features of sensory structures
  - Approaches to the study of sensing
- Survey of some of the human senses 166
  - Cutaneous (skin) senses
  - Kinesthetic (motion) sense
  - Vestibular sense (equilibrium)
  - Taste (gustatory) sense
  - Smell (olfactory) sense

Human vision: structure and function of the eye 172
- Anatomy of the visual apparatus 172
  - Structures auxiliary to the eye
  - The eye
- The visual process 178
  - The work of the auxiliary structures
  - The work of the optical lens system
  - The work of the retina
  - The higher visual centres
  - Some perceptual aspects of vision
  - Electrophysiology of the visual centres
- Eye diseases and visual disorders 197
  - The outer eye and auxiliary structures
  - The inner eye
  - Ocular injuries
  - Manifestations of systemic diseases
  - Visual disorders
  - Ophthalmological examination and corrective devices
  - Blindness

Human hearing and balance: structure and function of the ear 205
- Anatomy of the human ear 206
  - Outer ear
  - Tympanic membrane and middle ear
  - Inner ear
- The physiology of hearing 210
  - Transmission of sound waves through the outer and middle ear
  - Transmission of sound within the inner ear
  - Cochlear nerve and central auditory pathways
  - Hearing tests
- The physiology of balance: vestibular function 215
  - Detection of angular acceleration: dynamic equilibrium
  - Detection of linear acceleration: static equilibrium
- Ear diseases and hearing disorders 216
  - Outer ear
  - Middle ear
  - Inner ear

Bibliography 220

ANIMAL SENSORY RECEPTION

Nature and functions of sensory systems

CLASSIFICATION OF SENSORY SYSTEMS

According to location of receptors. In general, sense cells, or receptors, located superficially in an organism receive signals from outside the organism and are parts of the exteroceptive system. Receptors located inside the body receive signals from changes taking place inside the body and belong to the interoceptive system. On activation, sensory cells cause reactions appropriate to their location; they are said to respond with their local sign. For example, a decapitated frog reacts to stimulation of the skin by precisely directed limb movements aimed at wiping away the stimulus. Local sign in humans is expressed by a conscious awareness of the spot being stimulated, as when a person locates a thorn in the skin. This, however, is not true for vision, hearing, and smell, the sources of which are localized away from the body surface. Although some authorities believe that projection in space is learned, especially in humans, for most animals such ability seems to be innate. In many cases interoceptors stimulate channels that are never brought into consciousness; the presence of a local sign is thus shown only by the appropriateness of the resulting reactions. Internal pain is remarkable in that it is usually "misdirected" (referred) to the body surface in well-established patterns, according to its origin, a considerable help in medical diagnosis.

According to type of stimulus. More than one type of

energy applied to a sense cell can, if strong enough, generate a nerve impulse, which will be interpreted by the central nervous system (CNS) as a change in the specific energy to which the cell is sensitive and will cause the same results as if the appropriate stimulus were present. Thus, a specific reflex action can be brought about by natural stimulations, such as touch of the skin, as well as by electrical stimulation of the nerve fibres activated by such touch. Each type of sense cell thus causes a specific output reaction and a specific sensation, which is the modality perceived. In other words, if the optic nerve could be functionally connected to the ear and the acoustic nerve to the eye, lightning would be heard and thunder seen.

Selectivity with regard to specific energy changes comes about in diverse ways, the simplest of which is the localization of the sense cell in such a way that it is protected from unwanted stimuli and by the use of accessory structures that make it extremely sensitive to the wanted one. The sense cells in the eye, for instance, are protected from any but the most severe changes in mechanical pressure; at the same time, the eye's optical properties focus the incoming light on the layer of sense cells constituting the retina. The hair cells of the ear, which are very sensitive to rapid changes in air pressure because of the ear's structure, are also well protected from other mechanical disturbances by shock-absorbing fluid.

Another main factor that differentiates types of sense cells is the presence of specific receptor sites for reacting with the energy to which they are specifically sensitive. Certain cells, for example, can be specifically stimulated by a given substance and no other, at least in the small concentrations required for reaction. Cells with such narrow reaction ranges are rare, however; more often, each cell has a wider spectrum, as is the case of the photoreceptors of the eye with regard to colour. Photoreceptors comprise three types of cell, each with a definite optimum but reactive to well-overlapping band widths, thereby providing for a range of colour vision. In other cases, it is the threshold (the lowest energy level) to any given stimulus that varies in different cells; this variation provides for measurement of the intensity of the stimulus. In many cases, however, intensity is coded by the frequency of the nerve impulses each receptor sends to the central nervous system.

The actual amounts of energy that can be transformed into a nerve impulse are sometimes amazingly small. One or a few photons of light absorbed may suffice not only for reception and transformation into a nerve impulse in several optic fibres but also for visual perception.

Photoreceptors

Photoreceptors are sensitive to light changes. They contain photopigments for absorption of light. The variety of photopigments in different cells determines the number of colours that can be distinguished. It is interesting to note that in insects, among other animals, colour sensitivity is extended into the ultraviolet range, though it is short in the red range. Cells especially sensitive to infrared radiation are found in the remarkable pit organs of vipers, which enable the snake to locate warm-blooded prey from a distance even when it freezes into immobility.

In the skin of warm-blooded animals, nerve endings, with or without accessory structures, are present that react especially to warming or to cooling.

Taste and smell

Well-known organs of chemical reception are those of smell and of taste. Except in cases in which there is great specificity to one substance, as, for example, the sex attractant in insects, the spectrum of chemoreceptive cells is broad. The sense of taste was long thought to be mediated by narrow, separate fibres for acid, bitter, sweet, and sour sensations; this viewpoint is now being replaced by one in which the spectra are considerably wider. Frogs have been shown to have taste cells that react specifically to distilled water. Chemoreceptors are also present as interoceptors, a well-known example being the carotid body in certain vertebrates; this organ monitors oxygen pressure in the carotid artery, which supplies the brain with blood.

Mechanoreceptors are the most widespread type of sense receptor and the most varied with regard to localization, sensitivity, and type of nerve-impulse firing. There are numerous subdivisions of the mechanoreceptive sense, such as touch, pain, sound, gravity, and muscle tone. Examples in humans include the naked nerve endings in the cornea of the eye; the Pacinian corpuscles in the skin, with their multilayered sheathlike covers; and the hair cells in the inner ear. Impulse formation may continue for as long as stimulus lasts, thus giving a continuous (tonic) type of discharge, or be limited and proportional to the rate of change of the stimulus, thus producing an abrupt (phasic) discharge. A remarkable type of mechanoreceptor occurs in the elastic organs of crustacean legs; movement-sensitive cells fire for the time a joint moves in one direction, and others fire for the opposite movement.

Electroreception

Electroreception is known only in certain fishes. Electrosensitive cells are accompanied by an organ that sends out small or large voltage changes. The sense cells occur along the long axis of the fish, enabling it both to discover food objects in the surrounding water and to locate other fish. This system is a great aid for navigation in murky water (see ELECTRICITY AND MAGNETISM: *Bioelectric effects*).

Certain animals appear to be able to orient to environmental changes for which no specific sense cells are known. Among these, magnetism is the most outstanding example. In fish, magnetic fields may well be received by electroreceptors. In insects and birds, which seem to perceive magnetic fields, no special sense cells have been implicated. The wide variety of phenomena considered as extrasensory perception in man may be based on direct influence on central-nervous elements, thus bypassing sensory input channels.

EVOLUTION OF SENSORY SYSTEMS

Specific sensory abilities do not show a clear evolutionary progression, most likely because the development of any type of sense depends on many other factors in the total ecology of a given organism. Vision, for instance, is sometimes poor or absent in a species of a class in which other members have a highly developed visual system: examples include cave-dwelling species, relatives of sighted emergent species.

Mechanical stimuli are effective in all forms of life. Specialized organs, however, appear very early in animal evolution; such organs include gravity and light receptors in jellyfish. In more advanced members of the phyla Mollusca and Arthropoda, greatly developed sense organs occur, some of which show an amazingly close resemblance to vertebrate organs; *e.g., Octopus* eyes and semicircular canals (for equilibrium). There is always a close relationship between the presence of highly developed sense organs and a region of the central nervous system; the latter is needed to "process" the incoming information in order to abstract the cues of importance to a given animal. The fact that such elaborate systems exist does not exclude the possibility of much shorter and simpler pathways, which provide for more localized and quicker reactions; for instance, the blink reflex, caused by the sudden approach of an object to the eye, bypasses the visual cortex of the brain.

INTEGRATION OF SENSORY INFORMATION

Although sensory information must be coded into a flow of nerve impulses for transmittal over distances, interactions between adjacent sense cells and sensory neurons also occur. Nerve cells can influence each other by mutual connections that result in membrane potential changes (electrical differences) in one when the other is stimulated. Similar effects are often caused by nerve impulses. When a number of nerve cells (neurons) with adjacent receptive fields are activated, it is common to find that the ones receiving the strongest stimulation suppress the response of those that are stimulated less. This action leads to a sharper difference at boundaries of stimulated and nonstimulated areas; thus, contrasts can be enhanced by this process, known as lateral inhibition. Certain sense cells have the property of being active, usually at a low rate, when not stimulated. This activity can then be either increased or decreased by appropriate stimuli. It is by such means that neurons indicating visual movements in one direction or sounds changing from one frequency to another obtain their selectivity. Such elaborations can be performed at different levels of the central nervous system. In the higher

Lateral inhibition

mammals, for instance, the fibres forming the optic nerve are mainly of two types, with small visual fields, one in which light in the centre of the field excites while light in the surrounding field inhibits, and vice versa. In animals even as highly developed as the rabbit, more complex integration has taken place in the visual periphery, and optic fibres can indicate such features as the movement of oriented lines in specific directions, which in cats and monkeys does not seem to occur before units in the brain have sampled the incoming information.

From this and other information it is clear that the use the animal makes of its senses is highly correlated with the type of sensory integration taking place in the nervous system. The ways by which a given stimulus is analyzed are varied and as yet only partially understood. It is, however, possible to build models that can be of mutual benefit for engineering and sensory information processing. By feeding back part of the incoming signal to earlier steps in the information processing, stability is greatly enhanced both in organisms and in machines. (C.A.G.W.)

Mechanoreception

Sensitivity to mechanical stimuli is a common endowment among animals. In addition to mediating the sense of touch, mechanoreception is the function of a number of specialized sense organs, some found only in particular groups of animals. Thus, some mechanoreceptors act to inform the animal of changes in bodily posture, others help detect painful stimuli, and still others serve the sense of hearing (see below *Sound reception*).

Slight deformation of any mechanoreceptive nerve cell ending results in electrical changes, called receptor or generator potentials, at the outer surface of the cell; this, in turn, induces the appearance of impulses ("spikes") in the associated nerve fibre. Laboratory devices such as the cathode-ray oscilloscope are used to record and to observe these electrical events in the study of mechanoreceptors. Beyond this electrophysiological approach, mechanoreceptive functions are also investigated more indirectly—*i.e.*, on the basis of behavioral responses to mechanical stimuli. These responses include bodily movements (*e.g.*, locomotion), changes in respiration or heartbeat, glandular activity, skin-colour changes, and (in the case of man) verbal reports of mechanoreceptive sensations. The behavioral method sometimes is combined with partial or total surgical elimination of the sense organs involved. Not all the electrophysiologically effective mechanical stimuli evoke a behavioral response; the central nervous system (brain and spinal cord) acts to screen or to select nerve impulses from receptor neurons.

Man experiences sharp, localized pain as a result of stimulation of "pain spots" (probably free nerve endings) in the skin, and dull pain, usually difficult to localize, associated with inner organs. The sensory structures of pain spots in the skin differ from other receptors in that they respond to a wide range of harmful (noxious or nociceptive) stimuli. Excessive stimulation of any kind (*e.g.*, mechanical, thermal, or chemical) may produce the human experience of pain. Apart from eliciting this subjective feeling of pain, stimulation of pain receptors in the human skin is objectively characterized by such signs of emotional expression as weeping and by efforts to withdraw from the stimulus. The reflex withdrawal of his hand from a burning stimulus may begin even before the person becomes conscious of the pain sensation.

Experience of pain in animals

Judging from objective criteria, responses to painful stimuli also occur in nonhuman animals, but, of course, any subjective experience of pain sensation cannot be directly reported. Still, the question of painful experience among animals is of considerable interest because investigators (*e.g.*, medical researchers) are often obliged to subject laboratory animals to treatments that would elicit complaints of pain from a man. If a cat's tail is accidentally stepped on, the pitiful screeching and efforts to withdraw are so strikingly similar to human reactions that the observer is led to attribute the experience of pain to the animal. If one treads accidentally on an earthworm and observes the animal's apparently desperate struggles to get free, he might again be inclined to suppose that the worm feels pain. This sort of "mind reading," however, is inherently uncertain and may be grossly misleading.

The following observations illustrate some of the difficulties in making judgments of the inner experiences of creatures other than man. After the spinal cord of a fish has been cut, the front part of the animal may respond to gentle touch with lively movements, whereas the trunk, the part behind the incision, remains motionless. A light touch to the back part elicits slight movements of the body or fins behind the cut, but the head does not respond. A more intense ("painful") stimulus, however (for instance, pinching of the tail fin), makes the trunk perform "agonized" contortions, whereas the front part again remains calm. To attribute pain sensation to the "painfully" writhing (but neurally isolated) rear end of a fish would fly in the face of evidence that persons with similarly severed spinal cords report absolutely no feeling (pain, pressure, or whatever) below the point at which their cords were cut.

Aversive responses to noxious stimuli nevertheless have a major adaptive role in avoiding bodily injury. Without them, the animal may even become a predator against itself; bats and rats, for instance, chew on their own feet when their limbs are made insensitive by nerve cutting. Some insects normally show no signs of painful experience at all. A dragonfly, for example, may eat much of its own abdomen if its tail end is brought into the mouthparts. Removal of part of the abdomen of a honeybee does not stop the animal's feeding. If the head of a blow-fly (*Phormia*) is cut off, it nevertheless stretches its tubular feeding organ (proboscis) and begins to suck if its chemoreceptors (labellae) are brought in touch with a sugar solution; the ingested solution simply flows out at the severed neck.

At any rate, responsiveness to mechanical deformation is a basic property of living matter; even a one-celled organism such as an amoeba shows withdrawal responses to touch. The evolutionary course of mechanoreception in the development of such complex functions as gravity detection and sound-wave reception leaves much room for speculation and scholarly disagreement.

RECEPTION OF EXTERNAL MECHANICAL STIMULI

The sense of touch. Sensitivity to direct tactual stimulation—*i.e.*, to contact with relatively solid objects (tangoreception)—is found quite generally, from one-celled organisms up to and including man. Usually the whole body surface is tangoreceptive, except for parts covered by thick, rigid shells (as in mollusks). Mechanical contact locally deforms the body surface; receptors typically are touch spots or free nerve endings within the skin, often associated with such specialized structures as tactile hairs. The skin area served by one nerve fibre (or sensory unit) is called a receptive field, although such fields overlap considerably. Particularly sensitive, exposed body parts are sometimes called organs of touch—*e.g.*, the tentacles of the octopus, the beak of the sandpiper, the snout of the pig, or the human hand.

Stimulation of the human skin with a bristle reveals that touch (pressure) sensation is evoked only from certain spots. These pressure spots, especially those on hairless parts (*e.g.*, palm of the hand, or sole of the foot), are associated with specialized microscopic structures (corpuscles) in the skin. Pressure spots are most densely concentrated on the tip of the human tongue (about 200 of them per square centimetre, or 1,300 per square inch), roughly twice their concentration at the fingertip. A characteristic feature of many tactile sense organs is their rapid and complete adaptation (*i.e.*, temporary loss of sensitivity) when stimulated. Still, in man a distinction can be made between transient and more prolonged pressure sensations.

Relatively little research has been done with regard to the physiology of individual tangoreceptors in vertebrates. The Pacinian corpuscle of higher vertebrates, however, has been studied in isolation (see *Human sensory reception* below, for illustrations). These corpuscles, found under the skin, are scattered within the body, particularly around muscles and joints. Local pressure exerted at the surface or within the body causes deformation of parts of the corpuscle, a shift of chemical ions (*e.g.*, sodium,

potassium), and the appearance of a receptor potential at the nerve ending. This receptor potential, on reaching sufficient (threshold) strength, acts to generate a nerve impulse within the corpuscle. Among insects, movements of tactile hairs have been shown (sometimes specifically) to affect the receptor potential and the impulse frequency in the connected nerve fibre.

Many vertebrates and invertebrates can localize with some precision points of tactual stimulation at the body surface. People typically can still distinguish two sharpened pencil points, or similar pointed stimuli, when the points are separated by as little as about one millimetre (0.04 inch) at the tip of the tongue. (When moved closer together, the two points are perceived as one.) The human two-point threshold is about two millimetres at the finger tip, reaching six or seven centimetres (2.4–2.8 inches) at the skin of the back. Such tactual ability serves blind people when they read raised type (Braille) with their fingers. Closely related functions include the ability to distinguish between tactile stimuli that differ qualitatively; for example, between a rough and a smooth surface. This ability is even observable in the ciliate *Stylonychia* (a one-celled relative of *Paramecium*).

Reading Braille

Sensory contact with the ground below often informs animals about their spatial position. Nocturnal animals (for example, some eels) find shelter during the day by keeping as much of their skin as possible in contact with solid objects in the surroundings (thigmotaxis). Animals that live in running water usually maintain their position as they turn and swim head-on against the current (rheotaxis). Study of rheotaxic behaviour reveals that the sensory basis almost exclusively depends on visual or tactile stimuli (or both) arising from the animal's movements relative to the solid bottom or surroundings. The long antennae of many arthropods (*e.g.*, crayfish) and the lengthened tactile hairs (vibrissae) on the snouts of nocturnally active mammals (*e.g.*, cat, rat) serve in tactually sensing objects in the vicinity of the animal's body, extending and enriching the adaptive function of the sense of touch.

Lateral-line organs. *Mechanoreceptor function.* All of the primarily aquatic vertebrates—cyclostomes (*e.g.*, lampreys), fish, and amphibians—have in their outer skin (epidermis) special mechanoreceptors called lateralline organs. These organs are sensitive to minute, local water displacements, particularly those produced by other animals moving in the water. In this way, approaching organisms are detected and localized nearby before actual bodily contact takes place. Thus the lateral lines are said to function as receptors for touch at a distance, serving to perceive and locate prey, approaching enemies, or members of the animal's own species (*e.g.*, in sexual-display behaviour).

Each epidermal organ, called a sense-hillock or neuromast (Figure 1C), consists of a cluster of pear-shaped sensory cells surrounded by long, slender supporting cells. The sense hairs on top of the sensory cells project into a jellylike substance (the cupula) that bends in response to water displacement. The cupula stands freely in the surrounding water, grows continuously (*e.g.*, as a human fingernail), and wears away at the top. Sense organs of this type are distributed along definite lateral lines on the head and body of the animals (Figure 1A), developing in the outer layer of cells (ectoderm) of the embryo from a thickening called the lateral placode. From the central part of the same placode the sensory cells of inner-ear structures (the labyrinth) arise. The common embryologic origin and structural similarities of mature neuromasts and labyrinthine cell groups have led to the designation of all of these organs as the acoustico-lateralis system. The nerves to all the sense organs of the system arise from a common neural centre (called the acoustic tubercle in the wall of the brain's medulla oblongata). Among such amphibians as frogs, lateral-line organs and their neural connections disappear during the metamorphosis of tadpoles; as adults they no longer need to feed under water. The higher land-inhabiting vertebrates—reptiles, birds, and mammals—do not possess the lateral-line organs; only the deeply situated, labyrinthine sense organs persist.

The sensory cell of a neuromast bears one relatively long hair (kinocilium) and about 50 shorter ones (stereocilia). The kinocilium is inserted eccentrically on top of the sense cell; the stereocilia are arranged in parallel rows. In about half of the hair cells of a neuromast, the kinocilium is found on one (and the same) side of the cell; in the remaining hair cells it is found on the opposite side. In most cases these are cranial and caudal side, respectively. In the clawed frog (*Xenopus*), each group of hair cells in a neuromast connects to its own nerve fibre; hence there are two fibres per sense organ. The hair cells send a continuous series of neural impulses toward the acoustic tubercle in the absence of adequate external stimulation. A longitudinal water current along the toad's body surface, however, selectively increases or decreases the frequency of impulses from the cranial and caudal cells, depending on whether the flow is from head to tail or vice versa; current directed at right angles to such neuromasts has no effect. The impact of the moving water moves the cupula to deform the sensory hairs. Even minute cupula displacements of less than one thousandth of a millimetre are clearly effective in altering the impulses.

Detection of water currents

In *Xenopus*, as well as in other animals that have lateral-line organs, there are also some neuromasts with their hair cells asymmetrical at right angles to the head-tail axis. These add directional sensitivity so that other animals moving nearby in the water are well distinguished and localized. The postulated function of the lateral-line organs in the reception of low-frequency propagated pressure waves ("subsonic sound") has not been verified behaviorally. At very short distances, however, a vigorous low-frequency sound source stimulates the lateral-line system on the basis of acoustical near-field effects (water particle displacements), just as does any moving or approaching object.

Cyclostomes, many bony fishes, and all the aquatic amphibians studied have only superficial ("free") neuromasts of the kind described above. In the development of most fish, however, a number of structures called lateral-line canals (Figure 1B) are formed as a secondary specialization. They begin as grooves that develop in the epidermis along the main lateral lines; thus, a number of formerly free neuromasts are taken down to the bottom of each groove. The walls of the grooves then grow together above the neuromasts. Eventually the grown-together walls form canals under the epidermis, containing in their walls a series of canal neuromasts and a chain of openings to the outside (canal pores) along the lateral lines. The

From *Experientia* (1952)

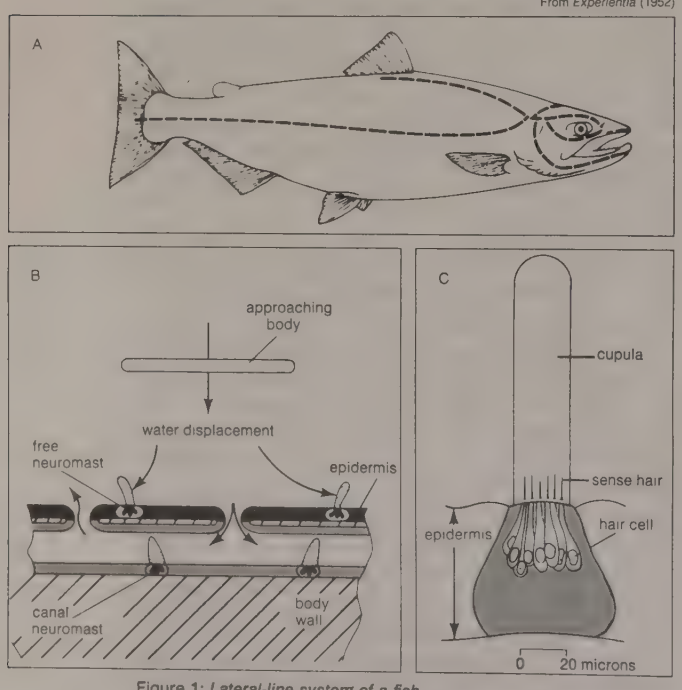


Figure 1: *Lateral-line system of a fish.*
(A) Bodily location of lateral lines; (B) longitudinal section of a canal; (C) superficial neuromast.

cupulae are changed in form, fitting the canal somewhat like swinging doors. The canal is filled with a watery fluid. Stimulation occurs essentially in the same way as with free neuromasts: local, external water displacement is transmitted via one or more canal pores to produce a local shift of the canal fluid to move cupulae. The sense cells in the canal neuromasts are polarized in the direction of the canal.

Canal specialization is particularly well developed in lively species of fish that swim more or less continuously and in bottom dwellers that live in running or tidal waters. Canalization has been interpreted as a case of adaptive evolution, serving to avoid the almost continuous, intense stimulation of free neuromasts by water flowing along the fish body during swimming or, in the case of relatively inactive bottom dwellers, by the external currents. These coarse water displacements probably mask subtly changing stimuli from detection by the lateral-line organs on the surface of the animal's body. Canal neuromasts are shielded in large degree from these masking currents.

Locating prey, predators, and sexual partners

The lateral-line organs function mainly in locating nearby moving prey, predators, and sexual partners. Usually these objects must be much closer than one length of the animal's body to be detected in this way; even intense stimuli are hardly ever detected beyond five body lengths away. Lateral-line function in rheotactic orientation against currents is restricted mainly to inhabitants of small currents such as mountain brooks, where marked differences of water-flow velocity affecting the fish body locally are likely to occur. Compared to their use of other sensory functions (*e.g.*, vision) the animals depend little on ability to sense extremely close, resting objects (obstacles) through the lateral lines. Obstacle detection of this kind does not arise from reflection of water waves; rather, the pattern of water displacement around the moving fish abruptly undergoes deformation at the near approach of an obstacle as the result of compression; the fish encounters a sudden rise in water resistance in the immediate vicinity of the obstruction. Nor do the lateral-line organs function to regulate or coordinate the animal's movements on the basis of the water flow or pressure variations along its body produced by swimming; neither do they serve for the reception of water-transmitted propagated sound waves (hearing).

Ampullary lateral-line organs (electroreceptors). Perhaps the most interesting specialization of the lateralline system is the formation in several groups of fish of deeply buried, single electrically sensitive organs. Such structures, for example, are found on the head of all the elasmobranchs (*e.g.*, sharks and rays), and are called ampullae of Lorenzini. Similar organs include those on the head of *Plotosus*, a marine bony fish (teleost); structures called mormyromasts in freshwater African fish (mormyrids) and in electric eels (gymnotids); what are named small pit organs of catfishes (silurids); and possible related organs in several other fish groups. These are known as ampullary lateral-line organs, and they have features in common. The sensory cells are withdrawn from the body surface, lack kinocilia, and have no mechanical contact with the surrounding water through a cupula. The latter attribute, indeed, is typical for all the acousticolateral end organs, except ampullary sense organs, in which the sense cells lie within the wall of a vesicle (or ampulla) that opens to the surface through a tubelike duct. Ampulla and duct are filled with a gelatinous substance that has excellent electrical conductivity.

Fish with ampullary sense organs are found to be remarkably sensitive to electrical stimuli—*i.e.*, minute, local potential differences in the surrounding water at their body surface. In behavioral experiments with sharks and rays, sensitivity to changes of 0.01 microvolt per centimetre (one microvolt = 1/1,000,000 of a volt) along the body surface has been found for the ampullae of Lorenzini. Similar, though somewhat higher, values have been recorded from the ampullary nerve fibres. A decrease in voltage at the opening of the ampulla causes an increase of the spontaneous nerve-impulse frequency; an increase in voltage at the opening produces the opposite response. Through their electrical sensitivity, such fish can detect and locate other organisms in darkness, in turbid water, or even when these organisms are hidden in the sand or in the mud of the bottom.

Sharks, rays, and most catfishes are able to detect electrical changes (biopotentials) emanating from other organisms. The freshwater mormyrids and eels, on the other hand, have special signal-emitting electric organs. They produce a series of weak electric shocks (up to a few volts), sometimes quite regularly and frequently; for example, about 300 shocks per second in the mormyrid fish *Gymnarchus*. In this way, a self-generated electric field is created in the immediate surroundings. Any appropriate object (for example, a prey animal with good conductivity in relation to fresh water) will cause a deformation of the electric field and can thus be detected in a radar-like manner through the sensitive ampullar electroreceptors.

Electrical "radar"

Some theorists suggest that initially mechanoreceptive lateral-line organs evolved into electroreceptors. At any rate, evidence of a certain double sensitivity—to mechanical and to electrical stimuli—has been observed in electrophysiological experiments with Lorenzinian ampullae. This double sensitivity has not been found, however, in behavioral experiments; alterations in behaviour indicate that ampullary lateral-line organs merely serve the animal as electroreceptors in adapting to the environment.

Other varieties of mechanoreception. *Surface waves.* Several species of animals living at or near the water surface use surface waves or ripples emanating from potential or struggling victims to locate their prey quickly: examples are the toad *Xenopus*, several fish species, and such insects as the back swimmer (*Notonecta*) and the water strider (*Gerris*). The whirligig beetle (*Gyrinus*) also uses surface ripples to avoid collisions with obstacles and companions. The sensory structures involved range from specialized tactile hair receptors (trichobothria) to internally located cells (proprioceptors) in movable body appendages and lateral-line organs.

Water and air currents. Special water-displacement receptors found in lobsters (*Homarus*) are most reminiscent of the lateral-line organs in vertebrates. Water-current receptors also enable several kinds of bottom-dwelling invertebrates to orient themselves (rheotaxis) in rivers and tidal currents. Many predators among these animals also respond chemically (see *Chemoreception* below), moving against the current (positive rheotaxis) until the prey is reached. In this way, for example, certain marine snails easily find their particular prey (sea anemones). Similarly among insects, the chemical "smell" of prey or of potential sex partners elicits a tendency to move against the wind (anemotaxis) until the source of the chemical stimulus is found. Several types of air-current receptors (true mechanoreceptors) on the heads of insects enhance such chemoreceptive behaviour. In flying locusts, an air current directed appropriately toward the head elicits compensatory reflex flight movements. The receptors involved (groups of hair sensilla on the head) mediate small corrections in the maintenance of straight flight; major guidance, however, derives from the insect's visual contact with the ground below.

Vibration reception. Adaptation and recovery occur most rapidly among touch receptors, and they tend to respond well to repeated stimulation, even of relatively high frequency. Thus, a person can feel whether an object is vibrating; above a threshold frequency of about 15 cycles per second (cps), discretely perceived tactual stimuli seem to fuse into a quite new and distinct vibratory sensation. The upper frequency limit of this vibration sense is found at several thousand cps among normal individuals, with sensitivity being maximal in the range of 200 cps (above a threshold amplitude of about 100 millimicrons). Just as pitch is discriminated in hearing, differences of about 12 to 15 percent in vibration frequencies can be distinguished by most people.

Vibration sensitivity is not limited to man; fish, for instance, also may respond to low-frequency water vibrations with tactile receptors. In addition, several kinds of animals have special vibration receptors. In some insects, a group of specialized structures (chordotonal sensilla) in the upper part of each tibial segment of the leg signal vibrations from the ground below. In the cockroach, the

Specialized vibration receptors

threshold amplitude for vibrational stimuli of this kind has been found to be less than 0.1 millimicron. Birds have special receptors (corpuscles of Herbst in the tibiotarsal bone of the leg) with which they can detect slight vibrations of the twig or branch on which they sit. Perhaps birds are alerted at night in this way to approaching predators; maximal sensitivity is at about 800 cps, and the threshold amplitude is close to 20 millimicrons. Spiders also use their vibration sense to locate prey in the web.

Generalized hydrostatic pressure. Several types of aquatic animals are sensitive to small changes of hydrostatic, or water, pressure. Among fish, this applies particularly to the order Ostariophysi (Cypriniformes), which includes about 70 percent of all the freshwater species of bony fishes. The swimbladder in these animals is connected with the labyrinth (sacculus) of the inner ear through a chain of movable tiny bones, or ossicles (weberian apparatus). Alterations in hydrostatic pressure change the volume of the swimbladder and thus stimulate the sacculus. These fish can easily be trained to respond selectively to minute increases or decreases in pressure (for example, to a few millimetres of water pressure), indicating that they have a most refined sense of water depth. All of these fish are so-called physostomes, which means that they have a swimbladder duct through which rapid gas exchange with the atmosphere can occur; many live in relatively shallow water. The hydrostatic-pressure sense can function to inform the animals about their distance from the surface or about the direction and velocity of their vertical displacement. It also appears that improvement and refinement of the sense of hearing arises through the swimbladder's connections via the weberian apparatus with the labyrinth.

Sensing of hydrostatic pressure

The sensitivity of several kinds of crustaceans to relatively small hydrostatic-pressure changes (as low as five to 10 centimetres [two to four inches] of water pressure) is most remarkable because these animals have no gas-filled cavity whatsoever. The mechanism by which the stimuli are detected remains a puzzling question, although information about changing water depth during tidal ebb and flow clearly would seem to have adaptive value.

RECEPTION OF INTERNAL MECHANICAL STIMULI

Some proprioceptors (internal receptors) for mechanical stimuli provide information about posture and movements of parts of the body relative to each other; others contribute to an undisturbed course of coordinated muscular actions (*e.g.*, in locomotion). Best known from studies of vertebrates and arthropods, some are tonic proprioceptors (serving to maintain muscle tone in posture); others are of the phasic type (serving movement); still others have a mixed phasic-tonic character. In principle, proprioceptors can be stimulated adequately by pressure or stretching during active movements of the animal (reafferent stimulation) as well as through passive external pushing and pulling (exafferent stimulation). One passive factor, particularly in land-inhabiting animals, is gravity as it acts on bodily tissues or organs. Proprioceptors thus not only serve reflex adjustments in posture and relatively automatic movements of parts of the body with respect to each other (as in driving an automobile) but they also provide gravitational information about the position of limbs or of the whole body in space. To the extent that they are gravity detectors, these sensory structures are properly called external receptors (exteroceptors instead of proprioceptors). For receptors that are diffusely located within the body, a clean distinction between proprioceptive and possible exteroceptive function (gravity reception) is experimentally practicable only under conditions of weightlessness, as in space travel.

Vertebrates. *Muscle spindles.* Well-known proprioceptors of all the four-limbed vertebrates studied are the muscle spindles occurring in the skeletal (striate) muscles; fish muscles show structurally simpler but functionally comparable receptors. Each muscle spindle in mammals consists of a few slender, specialized (intrafusal) muscle fibres that are surrounded by a sheath of connective tissue filled with lymph fluid (Figure 2). The muscle spindle itself is surrounded by and arranged parallel to the ordinary (extrafusal) muscle fibres. Each intrafusal fibre consists of contractile (motor) parts at both ends and a noncontractile sensory midsection that serves as a receptor for stretch (changes of length and tension). There is double (primary and secondary) sensory innervation in mammals, but the secondary endings are lacking in lower vertebrates. Even when the animal is at rest, both types of endings are active (under the tension of normal muscle tonus). Additional stretch (lengthening) of the intrafusal midsection increases the nerve impulse frequency, and relaxation (shortening) causes a decrease. The primary (phasic-tonic) ending responds quickly; responses of the secondary (tonic) endings are slower.

The length of the muscle spindle as a whole varies with the contraction phase and the length of the muscle to which it belongs. The length of the sensory midsection, however, may change more or less independently because its motor nerve endings function apart from the innervation of the extrafusal muscle fibres. Thus the ratio of extrafusal–intrafusal contraction determines whether or not a change of length in the midsection will occur during muscle activity. There are reasons to suppose that midsection stretch remains more or less unchanged during self-initiated ("voluntary") movements; reafferent stimulation of muscle spindles would be avoided in this way. But as soon as an unexpected (exafferent) stretch of a muscle occurs—for example, when a leg pushes against an obstacle during locomotion—the midsections stretch to produce an increase of impulse frequency. This neural activity elicits a compensatory reflex contraction of the stretched muscle, as in the knee jerk during medical examinations: a blow beneath the kneecap causes stretch of a thigh muscle, stimulation of its muscle spindles, and a compensatory jerking contraction of the same muscle.

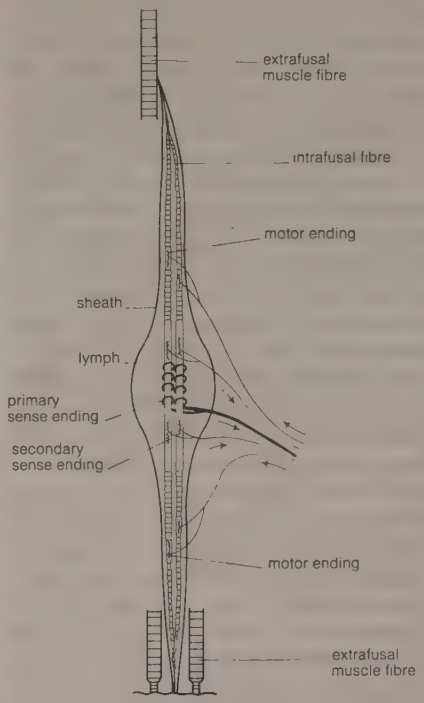


Figure 2: Mammalian muscle spindle.

Knee jerk

Tendon organs. Branched nerve endings on vertebrate tendons (not far from their point of attachment to muscle) also respond to stretch; however, they are decidedly less sensitive than are muscle spindles. These tendon organs produce no impulses under the stretch of normal, resting muscle tonus. Neither is there a mechanism preventing reafferent stimulation of tendon organs, nor does it make any difference whether the stretch is brought about by active muscle contraction or passively following external influence. In both cases tendon receptors respond according to the intensity of the stretch; their response causes relaxation of the attached muscle and may serve (among other functions) to prevent anatomical damage.

Human awareness of posture and movement of parts of the body with respect to each other (kinesthetic sensations) is attributable neither to muscle spindles nor to tendon organs. The sensations are based on stimulation of sensory nerve endings of various types at the joint capsules and of stretch receptors in the skin. There are also mechanoreceptors in the walls of some blood vessels (*e.g.*, in the aorta and the carotid sinus); these are sensitive to blood-pressure changes and play a regulatory role in the circulatory system.

Invertebrates. Among invertebrates, the arthropods exhibit the most readily distinguished proprioceptors, called muscle-receptor organs and chordotonal proprioceptors. Both types of structure occur in crustaceans as well as in insects. Adequate stimuli are variations in length and tension (stretch).

Muscle receptor organs. Although they structurally and functionally resemble the muscle spindles of vertebrates, arthropod muscle receptor organs are always situated outside of the muscles proper. Numerous branches of multipolar primary nerve cells are connected with the noncontractile midsection of specialized muscle fibres, both ends of which are contractile and have an efferent (motor) innervation. In crustaceans, the muscle receptor organ contains two elements: a slowly contracting, nonadapting tonic fibre and a quickly contracting, rapidly adapting phasic element.

Chordotonal proprioceptors. Widely distributed among arthropods, chordotonal receptor organs are thin, elastic, innervated strands of connective tissue, stretched between adjacent segments of the body or of leg joints. The sensory endings of a few bipolar primary nerve cells, each provided with a spiny sensillum (scolopidium), are attached to the strand. Chordotonal proprioceptor organs generate neural impulses that show them to contain both phasic movement receptors and tonic pressure receptors; sometimes two varieties of each. Thus there are receptors that selectively respond only during flexion, only in the flexed position, only during stretch, or only in the stretched state of the given strand. Several kinds of insects, apart from their clearly proprioceptive-chordotonal functions, have other chordotonal elements that serve as typical exteroceptors. Sense organs of this type (tympanic and subgenual organs in legs, organs of Johnston in the antennae) may function in the reception of sound waves, of vibrations in the ground, or of other external mechanical stimuli. Many insects also have a special type of chordotonal-proprioceptor structure (campaniform sensilla) not found in crustaceans. Sensory endings of primary nerve cells are connected with thin, dome-shaped (campaniform) spots on the exoskeleton. These campaniform sensilla respond to external stimuli such as local tensions and deformations of the body surface. They function in the regulation of such movements as the beating of wings in locusts. Similarly functioning proprioceptors (lyriform organs) are also observed among spiders.

In insects, body posture and movements of individual body parts with respect to each other can be detected through groups of external tactile hairs implanted near the joints between adjacent skeletal elements. Some function as rotation receptors or exteroceptors to detect the direction of gravity.

Among other invertebrates, the cephalopod *Octopus* clearly exhibits proprioceptive abilities, though specific receptors have not yet been identified. These animals, however, seem unable to integrate proprioceptive data in the central nervous system with other sensory information in learning. Thus an octopus readily can be taught to discriminate between two small cylindrical objects (both provided with longitudinal ribs) if the ribs on one of them are somewhat coarser than those on the other. But the animal cannot learn to distinguish between cylinders of the same size if the ribs are equally coarse and if they are longitudinal on one and transverse in the other; nor can it learn to discriminate between small objects of different form or different weight. This indicates that an octopus cannot learn any discrimination that depends on sensory information about the position of the arms and suckers making contact.

Octopus learning

MAINTENANCE OF EQUILIBRIUM

Active maintenance of equilibrium during bodily movement (*e.g.*, in locomotion) requires appropriate sensory functions. Although many animals usually maintain their bodies with the long axis horizontal (backside up), man being a notable exception, there are frequent departures from the usual position. A fish may dive steeply downward and a man may alter his normal orientation by lying down at full length. In no case, however, need there be any loss of equilibrium. Every deviation means an equilibrium disturbance and evokes compensatory reflex movements, not only a deviation from the usual position as in most laboratory experiments.

Maintenance of equilibrium is based upon contact of the animal with the external world; several sensory systems may play a role in this context. When an animal moves over a solid surface, tactile stimuli usually predominate as cues. It has been noted above how proprioceptors in vertebrates and arthropods can also contribute to spatial orientation; bodily tissues under gravity weigh vertically down and stimulate internal mechanoreceptors in a way that depends on, and varies with, the animal's spatial position. When they are out of contact with the ground, many animals orient themselves in space by keeping their back (dorsal) side turned up toward the light. Visual cues also can serve equilibration; for example, through compensatory body movements (optomotor reflexes) brought about by the shifts of the image of the environment over the retina of the eye. For the receptors mentioned thus far, however, equilibration is not the unique function. There are other sensory structures that are genuine organs of equilibrium in that they primarily and exclusively serve orientation of posture and movement in space.

Secondary cues

Gravity receptors. Because of the constancy of its magnitude and direction, gravity is most suitable in providing animals with cues to their position in space. The sense organs involved (statoreceptors) usually have the structure of a statocyst, a fluid-filled vesicle containing one or more sandy or stonelike elements (statoliths). Sensory cells in the wall of the vesicle have hairs that are in contact with the statolith, which always weighs vertically down. Hence, depending on the animal's position, different sense cells will be stimulated in statocysts with loose statoliths (Figure 3A); or the same sense cells will be stimulated in different ways in statocysts with a statolith loosely fixed to the sense hairs (Figure 3B).

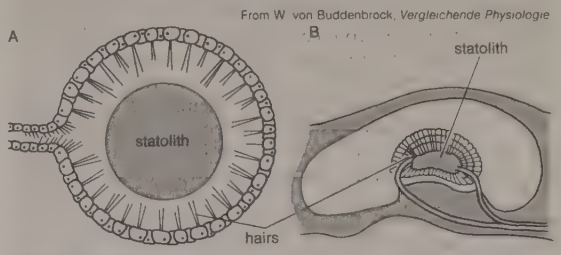


Figure 3: *Statocyst gravity receptors.*
(A) With a free-moving statolith, as in a mollusk (scallop), and (B) with statolith loosely fixed to hair cells, as in a crustacean (opossum shrimp).

Statocysts are found in representatives of all of the major groups of invertebrates: jellyfish, sandworms, higher crustaceans, some sea cucumbers, free-swimming tunicate larvae, and all the mollusks studied thus far. Analogous receptors that occur generally in vertebrates are the ear's utriculus and probably (to a degree) also two other otolith organs (sacculus and lagena) of the ear (labyrinth). Statocysts (including vertebrate labyrinthine statoreceptors) develop embryologically from local invaginations of the body surface. In primitive evolutionary forms, the interior of the statocyst is in open communication with the surrounding sea and thus is filled with water; statoliths usually are sand particles taken up from outside. In a few animal groups, this developmental stage is only found during the larval phase, the initial opening to the exterior being closed in the adult animal. In more advanced forms, the liquid content (statolymph) and the statoliths are produced by cells in the wall of the organ. This specialized

type of closed statocyst is found in many snails, in all the cephalopods such as the squid (except *Nautilus*), and in the vast majority of vertebrates (from bony fishes up to and including mammals).

Statocyst function may be studied by observing compensatory reflexes under experimental conditions. When the position of a laboratory animal is appropriately changed, movements of such body parts as the eyes, head, and limbs can be observed. Such movements tend to counteract the imposed change and to restore or to maintain the original position. Evidence of statoreceptor function is provided if these reflexes are abolished after surgical elimination of both statocysts. Many animals exhibit locomotion that is gravitationally directed vertically down or up (positive or negative geotaxis, respectively). Geotactic behaviour may be experimentally altered by whirling the animal in a centrifuge to change the direction and to increase the intensity of the force exerted on the sensory hairs by the statoliths. Molting crustaceans shed the contents of their statocysts along with their exoskeleton. If such an animal is placed in clean water containing iron filings, it takes up new iron statoliths instead of the usual sand grains. By moving a magnet to vary the direction of the force exerted by the metal statoliths, the animal can be made to adopt any resting position, even to stay upside down. Statoliths can be washed out of the open statocysts of a shrimp without damaging the sensory hairs. When the hairs are pushed in different directions with a fine water jet, the shrimp exhibits compensatory reflexes. In this way, it has been shown that each statocyst signals a change of position around the animal's long axis; the same reaction is found to occur after removal of the statocyst on one side only. Electrical impulses in the statocyst nerve can be recorded while the animal is in different spatial positions, or during experimental deflection of the sensory hairs. Such experiments reveal that both vertebrates and decapod crustaceans (*e.g.*, shrimp) exhibit spontaneous and statolith-induced neural activity in the lining (epithelium) of the gravity receptor.

Geotactic behaviour

Spontaneous activity. The sensory epithelium of a statocyst is spontaneously active, initiating a continuing series of impulses directed toward the central nervous system (even when the statoliths are experimentally removed from the statocyst). This resting frequency of neural activity is fairly constant and completely independent of the animal's position in space. In vertebrates and in crustaceans, spontaneous activity of the left statocyst affects the central nervous system to produce a tendency of the animal to roll to the right about its long axis; spontaneous activity of the right statocyst prompts a tendency to roll to the left. Normally, these rolling tendencies neutralize each other in the central nervous system, not becoming manifest unless the statocyst on one side of the body is functionally eliminated by complete surgical removal, by destruction of its sensory epithelium, or by cutting its nerve. This intervention permits the influence of the spontaneous activity generated in the remaining statocyst to be felt, and the animal tends to roll toward the operated side. Unilateral (one-sided) removal of the statoliths alone, however, does not produce such an effect so long as the sensory cells in the epithelium remain intact. The rolling tendency of a unilaterally operated animal usually diminishes little by little in the course of hours or days, until it finally disappears completely. If the remaining statocyst is then removed, rolling occurs again, but this time to the other (last operated) side. This tendency also diminishes and disappears with time. Apparently the unbalancing effect of the spontaneous influx from a statocyst is gradually counteracted in some unknown way by the central nervous system.

Statolith influences. Vertebrates and crustaceans have statoliths that are loosely connected to the sensory hairs by a sticky substance. With such a mechanical arrangement, the statolith stimulates the sensory cells by parallel (shearing) motion rather than by pressure or pull at right angles to the epithelium. The effects are demonstrable in experiments with fish, based on the dorsal-light orientation noted above. In a laboratory darkroom, if light shines at a fish from one side, the animal assumes an oblique position. While the fish tends to turn on its side (with its back side to the light), gravity tends to keep it vertical; the oblique position is the result. In a whirling centrifuge, the pressure exerted by the statoliths may be increased. When this is done, the fish rights itself almost precisely to the degree that the shearing force exerted by the statoliths is held constant.

Otolith organs in vertebrates

In vertebrates, statoreception is localized in the head within the labyrinth, particularly within the utriculus, one of the three statolith (or otolith) organs (Figure 4). The statolith is surrounded by a gelatinous substance akin to the cupula of the lateral-line organs. In most higher vertebrates, the head moves rather flexibly because it is not rigidly connected to the trunk. Thus information coming from the utriculi has to be neurally integrated centrally with impulses from proprioceptors that signal the position of the head with respect to the limbs and trunk (for example, neck receptors), if the animal is to orient its head and body appropriately in space.

The roles played by the remaining otolith organs of the labyrinth (sacculus and lagena) in statoreception remain unclear. Their sensory epitheliums (maculae) are roughly at right angles to each other and to that of the utriculus. In view of their arrangement, it was once supposed that the three otolith organs of the labyrinth would serve to detect position in three spatial planes (indeed, the three semicircular canals do serve to detect rotation in different planes). It has been found, however, that the sacculus and the lagena (as far as it is present) can be put out of function bilaterally in representatives of all the classes of vertebrates without causing overt equilibration disturbances. On the other hand, some secondary statoreceptor function has been demonstrated for these otolith organs in all the animals from fish up to and including man.

In the special case of flatfishes (*e.g.*, halibut, sole, flounder), the normal upright position in the juvenile stage changes to one of swimming and lying on one side as an adult. The eye from that side migrates to the upper surface; but the situation of both labyrinths remains unchanged. Hence, the originally horizontal maculae of the utriculi are now oriented vertically. In these fish, the sacculi (usually the major organs of hearing in bony fishes) indeed may be shown to serve as statoreceptors. At any rate, the same otolith organ may function in one fish species as an organ of hearing and in another as a gravity receptor; clearly, both functions depend on basically identical mechanical stimulation.

As receptors belonging to the acousticolateralis system, the otolith organs of vertebrates have hair cells of the same type that is found in lateral-line neuromasts. Under the electron microscope, the sensory hair cells show a pattern of polarization (arrows in Figure 4) throughout the macula, indicating the directions in which the shearing otolith should have an activating or an inhibiting influence. Results of physiological investigations thus far performed agree well with these deductions.

Statocyst research among crustaceans

Among the invertebrates, most statocyst research has been done with such decapod crustaceans as lobsters.

From Cold Spring Harbor Symposia on Quantitative Biology (1965)

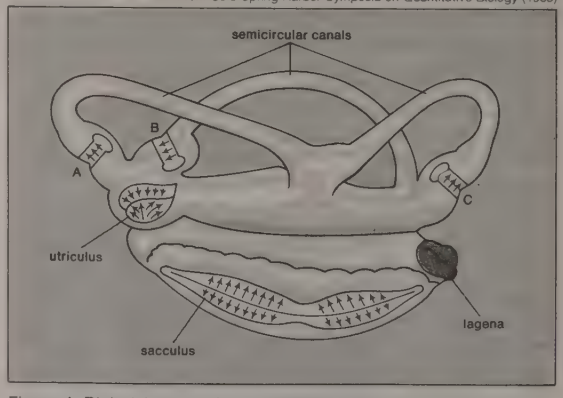


Figure 4: Right labyrinth of a codlike fish called the burbot, seen from above. Ampullae of anterior, horizontal, and posterior semicircular canals are lettered A, B, and C, respectively. Arrows show direction of hair-cell polarization.

The working mechanism of their statocysts conforms with the physiological principles of vertebrate statoreception discussed above. The results of electrophysiological investigations support the conclusions drawn from behavioral observations. In some crustacean statocysts (for example, in the lobster, *Homarus*), special statoreceptors are found that signal the same bodily position differently, depending on the direction of movement through which it was reached (hysteresis effect). The part played by the statocyst in equilibration has been investigated in several other invertebrate groups, among them jellyfish, sandworms, and such mollusks as scallops, common snails, sea hare, and octopus. Each sensory cell from the vertical macula in a statocyst of the octopus bears up to 200 kinocilia, and all the cilia of each cell are polarized in the same direction. On the macula as a whole, there is a radiating polarization pattern, the activating direction pointing everywhere from the centre to the margin. Compensatory eye reflexes resulting from tilting the animal head down or head up around a transverse axis reveal a hysteresis effect. After unilateral-statocyst removal, mollusks do not tend to roll toward the operated side (as do vertebrates and crustaceans) but toward the side of the remaining statocyst.

The almost complete absence of statocysts in insects is remarkable in view of evidence that many of them have a high degree of sensitivity to the direction of gravity. Receptors involved are specialized tufts of tactile hairs at the external body surface; in the honeybee, such groups of hairs are notably found between head and thorax and between thorax and abdomen. The adaptive function of these static (gravity) receptors becomes manifest in the honeybee "dance language" performed on a vertical comb in the hive. The angle between the dancing bee and the perpendicular seems to direct other bees to sources of nectar and pollen.

Rotation receptors. In addition to having tonic statoreceptors (signalling position with respect to gravity), several groups of animals have purely phasic rotation receptors that respond only to angular acceleration or deceleration, as produced on a turntable. Vertebrates, cephalopods (*e.g.*, squid), and decapod crustaceans (*e.g.*, lobsters) have special rotation receptors at the inner surface of the fluid-filled organ of equilibrium (labyrinth or statocyst). This fluid lags inertially with respect to the wall of the organ at the onset and arrest of every rotation. Among crustaceans, such as crabs or lobsters, the rotation receptors incorporate relatively long, delicate hairs that extend more or less at right angles to the wall freely into the statocyst fluid. The hairs respond quickly to fluid motion, swaying around their point of attachment and returning slowly through their elasticity to resting position. Their stimulation causes compensatory reflexes of the eyestalks or of the whole animal.

Eyestalk reflexes can be readily observed when a blinded, legless crab is rotated while flat on a turntable. These reflex movements are called nystagmus. At the onset of rotation to the right, both eyestalks move at about the rotation rate to the left by way of compensation until they reach their maximal deviation. In most cases, one or more jerky movements of the eyestalks in the opposite direction are observed per rotation during the initial period (quick, restoring nystagmus phases). In general, however, the eyestalks remain deviated opposite to the direction of rotation for several revolutions of the turntable. During prolonged constant-velocity rotation, the crab's eyestalks return to their symmetrical position; at this point, inertial lag in the statolymph is reduced to the degree that the fluid finally rotates together with the statocyst wall. Sudden arrest of the turntable under these circumstances causes afternystagmus: the eyestalks move promptly to the right (at about the same velocity as they moved to the left at the onset of rotation) until their maximal deviation toward this side is reached. After a quick jerk in the opposite direction, the eyes continue their slow movement to the right, and in this way as many as three or more after nystagmus jerks may occur with decreasing intensity. Such aftereffects may last many seconds, but finally the eyestalks return slowly to their symmetrical position. All of these nystagmic effects from such horizontal rotation are abolished in a blinded crab, however, after bilateral elimination of the long, delicate statocyst hairs by their denervation or by cauterization of the hair bases.

Rotation of a blinded crab

In vertebrates, rotation reception occurs within the labyrinth. Each labyrinth has three semicircular canals arranged in planes at right angles to each other (Figure 4); the canals communicate with the utriculus. One end of each canal is widened into an ampulla, and the sensory cells (hair cells) are arranged in a row on a ridge (crista) of the ampullar wall. The crista is oriented at right angles to the plane of the canal, and the extended hairs of its sensory cells are imbedded in a jellylike cupula that reaches to the opposite wall of the ampulla. Endolymph displacement through a canal makes the cupula move aside, as if it were a swinging door. In vertebrates, the inertial lag of the endolymph at the onset of rotation is very brief, the fluid catching up with the angular velocity of the labyrinth within a fraction of a second. An ampulla with its crista and cupula is reminiscent of a lateral-line canal neuromast, except that all of the hair cells of a crista are polarized in the same direction. In the cristae belonging to the vertical semicircular canals the kinocilium is implanted at the side facing the canal; in the horizontal cristae all of the sensory cells are polarized toward the opposite side (facing the utriculus). This structural arrangement is in keeping with differences between the vertical and the horizontal canals observed in behavioral and electrophysiological experiments.

A turn of the animal's head around the vertical axis to the left increases the neural-impulse frequency (activation) in the left horizontal crista; a turn of the head to the right causes a frequency decrease (inhibition). Opposite effects occur at the same time in the right horizontal crista. Recordings from the different crista nerves, while the animal is being rotated successively around all three major body axes, show the horizontal crista to respond only to rotation of the animal around its vertical axis; the vertical cristae, however, respond to rotation about all three axes. Behavioral data (compensatory eye reflexes) provide similar results, except that the eyes fail to exhibit an observable response associated with the vertical semicircular canals during rotation around the vertical axis. Stimulation of the vertical cristae under these circumstances gives rise to the simultaneous contraction of antagonistic pairs of eye muscles; hence the absence of a compensatory eye rotation.

In decapod crustaceans, particularly crabs, the statocyst is anything but a simple spherical vesicle; it has a very complicated shape with several curved invaginations and projections. In a small corner in the lowest (most ventral) part of the crab statocyst, a cluster of minuscule sand particles (statoliths) is found in contact with specialized (hooked) hairs. Apart from these hook-hair gravity receptors, there is a single, slightly curved row of relatively straight "thread" hairs atop an oval invagination in the middle of the lower statocyst wall. These hairs are the rotation receptors described above in the blinded crab revolving on a turntable. Bilateral elimination of the thread hairs alters the reflexes of the eyestalks. Instead of reacting immediately, at the very onset of rotation, in the absence of thread hairs on both sides, the eyes initially maintain their symmetrical position. They start their compensatory movement only after rotation has begun or at the end of rotation after the animal has reached a new, steady position. Furthermore, the velocity of this compensatory eye movement seems to be independent of the rate of angular acceleration or deceleration. The delayed nature of the response suggests that loss of rotation sensitivity about horizontal axes results from thread-hair elimination.

That the thread hairs are indeed responsive to rotation about all three major body axes is supported by a number of observations. Bilateral elimination of the impulses from the statolith hairs (position receptors) by selective nerve cutting, for example, does not affect the animal's response to rotation around the vertical axis. Despite their loss of impulses from position receptors, crabs subjected to angular acceleration or deceleration about either horizontal axis exhibit the normal compensatory eyestalk reflexes at the very onset of rotation. When a new (inclined) position of the animal is maintained, the eyestalks again become

Eyestalk reflexes

symmetrical, although complete return to symmetry may require several minutes. On the other hand, when both thread hairs and statolith hairs are eliminated, all such rotation and position reflexes of the eyestalks and related aftereffects are abolished. After unilateral elimination of the thread hairs or removal of one entire statocyst in a blinded crab, both eyes still react to rotation around the vertical axis in both directions. When electrical recordings are made of the activity of the primary sensory neurons innervating the thread hairs, similar results are obtained, the receptors responding only to angular acceleration and deceleration. They are spontaneously active, and the neural response to rotation that is superimposed upon the spontaneous background consists of a coded sequence of impulse-frequency increases and decreases. The same reception unit responds to acceleration about all three major axes.

The statocysts of cephalopods (nautilus, squid, octopus) rival the complexity of crab statocysts. In addition to the perpendicular macula with its statolith (for gravity reception), the octopus has three cristae (containing many hair cells with two-directionally polarized kinocilia) arranged approximately at right angles to each other. Rotation (turntable) experiments and surgical removal of statocyst receptors have shown that the octopus cristae function as rotation receptors. Nystagmus and afternystagmus persist almost unchanged after unilateral statocyst removal, but they are completely abolished after the additional removal of the second statocyst in a blinded octopus. In the cuttlefish (*Sepia*), the statocyst is structurally even more complicated; besides three cristae, it has three maculae (statolith organs) also arranged in different planes.

Rotation receptors of a different type are found in some groups of insects. Dragonflies (for example, *Aeshna*) have external hair receptors between the head and thorax. If a gust of wind turns the animal around its long axis during flight, the relatively heavy head lags with respect to the thorax. The resulting stimulation of the hair receptors in the neck region elicits compensatory flight reflexes and restores the insect to a normal position. These receptors do not respond to static head displacements. In the Diptera (true flies), the posterior knobbed "wings" (halteres) serve as flight stabilizing rotation receptors. During flight, the halteres beat in a vertical plane, synchronously with the forewings. Rotational instability is gyroscopically counteracted by the beating action. Receptors are campaniform sensilla at the base of the haltere. (S.Di.)

Thermoreception

Thermoreception is a process in which different levels of heat energy (temperatures) are detected by living things. Temperature has a profound influence upon living organisms. Active life among animals is feasible only within a narrow range of body temperatures, the extremes being about 0° C and 45° C. On the Fahrenheit scale the same range is 32° F and 113° F. Limitations depend on the freezing of tissues at the lower temperature and on the chemical alteration of body proteins at the higher end of the range. Within these limits the metabolic rate of the animal tends to increase and decrease in parallel with its body temperature.

Body temperature and metabolism among more highly evolved animals (*e.g.*, birds and mammals) are relatively independent of direct thermal influences from the environment. Such animals can maintain considerable inner physiological stability under changing environmental conditions and are adaptable to substantial geographic and seasonal temperature fluctuations. A polar bear, for example, can function both in a zoo during summer heat and on an ice floe in frigid Arctic waters. This kind of flexibility is supported by the function of specific sensory structures called thermoreceptors (or thermosensors), which enable the animal to detect thermal changes and to adjust accordingly.

Temperature of the body directly reflects that of the environment among cold-blooded (poikilothermic) animals, such as insects, snakes, and lizards. These creatures maintain safe body temperatures mainly by moving into locations of favourable temperature (*e.g.*, in the shade of a desert rock). Warm-blooded (homoiothermic) organisms, such as the polar bear, normally keep practically constant body temperature, independent of environment. Homoiothermic animals, including man, are able to control their body temperature not only by moving into favourable environments but also through the internal regulatory (autonomic) effects of the nervous system on heat production and loss. Such autonomic adjustments depend on lower brain centres; the behavioral (movement) responses require the function of the brain's outer layers (the cerebral cortex).

Cold-blooded and warm-blooded animals

A variety of behavioral responses is elicited through stimulation of thermoreceptors, including changes in body posture that help regulate heat loss and the huddling together of a group of animals in cold weather. In some species, thermoreceptors are also involved in food location and sexual activities. Bloodsucking insects, such as mosquitoes, are attracted by thermal (infrared) radiations of warm-blooded hosts; such snakes as pit vipers can locate warm prey at considerable distance by means of extremely sensitive infrared receptors. Man has achieved the widest range of adaptability to extremes in temperature, since his technology allows him to protect himself under a considerable variety of thermal conditions on earth and even in outer space.

Perceptual aspects of thermoreception are found in evidence that man and other animals have conscious temperature sensations and emotional experiences of thermal comfort and discomfort (see below *Human sensory reception*). The effects of temperature on productive efficiency and behaviour (*e.g.*, on one's ability to think) have led to the installation of heat-regulating equipment in homes, public buildings, factories, and similar shelters for people, livestock, and other animals.

Thermoreception can be studied in different ways: (1) on the basis of reports of temperature sensations and thermal comfort by human subjects; (2) through observations of behavioral responses to variations in temperature by all kinds of animals; (3) by the measurement of compensatory autonomic responses (*e.g.*, sweating or panting) to thermal disturbances in the environment; and (4) by recording electrical impulses generated in the nerve fibres of thermoreceptors in laboratory animals and human subjects.

GENERAL PROPERTIES OF THERMORECEPTORS

The concept of thermoreceptors derives from studies of human sensory physiology, in particular from the discovery reported in 1882 that thermal sensations are associated with stimulation of localized sensory spots in the skin. Detailed investigations reveal a distinction between hot spots and cold spots; that is, specific places in the human skin that are selectively sensitive to warm stimuli or to cold. To this extent the different thermoreceptors exhibit sensory specificity. Modern neurophysiological methods show thermoreceptors also to be biophysically specific, in that they include nerve endings that are excited only by or primarily by thermal stimuli.

Early investigations of thermoreception

Extending far beyond the context of conscious temperature sensation as reported by humans, the biophysical definition holds for any thermoreceptive structure. Clearly, electrical responses from thermoreceptors are observable whether conscious sensations are reported by the animal (as in the case of a person) or whether they are not (as in the case of a laboratory rat). Although they are closely related, the concepts of sensory and biophysical specificity are not identical, the criterion being the quality of inner experience (sensation) in the first case and the quality of the neurally effective stimulus in the second. To make the distinction clear, a receptor that is neurally excited by cooling as well as by the application of a chemical (*e.g.*, menthol) might be classified only as a specific (cold) thermoreceptor in terms of human sensation; biophysically, however, it manifestly is a chemoreceptor as well (see below *Chemoreception*).

Most of the modern understanding of thermoreceptors is based on biophysical (electrophysiological) investigations. This approach, introduced in 1936 for recording the electrical signals from single thermosensitive nerve fibres in

the tongue of the cat, had been applied by 1960 to similar recordings from single thermoreceptors in the skin of human subjects. Such investigations are made by dissecting single nerve fibres under the microscope and placing them on electrodes or by inserting very fine wires (*e.g.*, tungsten microelectrodes) directly into the intact nerve or receptor. As in the case of other sensory nerve fibres, the electrical signals generated by the activity of thermoreceptors are brief impulses of about one millisecond duration and roughly constant amplitude. They follow in a more or less regular sequence, modulations (changes) in the frequency of which reflect differences in the intensity of the stimulus. (Frequency modulation is widely applied in such devices as radios for information processing.) Sensory structures are called specific thermoreceptors if they respond biophysically to temperature stimuli yet are practically insensitive to such other kinds of stimulation as mechanical pressure.

The general properties of thermoreceptors in the external parts of the body are found to be similar for any species of animals investigated. Thermoreceptors can be divided into well-defined classes as cold and as warm receptors. At constant temperatures (within an appropriate range), cold receptors are continuously active electrically, the frequency of the steady discharge (static response) depending on temperature. In most cases the static activity reaches a maximum at temperatures between 20° and 30° C (68° and 86° F). On sudden cooling to a lower temperature level, the cold receptors respond with a transient increase in frequency (dynamic response); if the lower temperature is maintained, the frequency drops to a level of static discharge in adaptation. When the receptor is warmed up again, a transient decrease in electrical activity is seen, after which the frequency rises again and finally adapts to the initial static value. Warm receptors are also continuously active at constant temperatures, with a maximum at 41° to 46° C (about 106° to 115° F). On sudden temperature changes, warm receptors respond in the opposite direction from that of cold receptors, temporarily overshooting adaptation frequency on warming and showing transient inhibition on cooling. Thermoreceptors are thus selectively sensitive to specific ranges of temperature as well as to rate of temperature change.

Cold receptors and warm receptors

Some receptor cells in the skin of fishes and amphibians respond both to mechanical and to thermal stimulation. In the skin of cat, monkey, and man, receptors have been found that are excited both by mechanical stimuli and by cooling. It seems, however, that these nerve endings are primarily mechanoreceptors (see above *Mechanoreception*), their sensitivity to cooling being much lower than that of specific cold receptors.

THERMORECEPTORS IN INVERTEBRATES

Insects placed on a surface that provides a temperature gradient (warmer at one end and cooler at the other) often congregate in a narrow band of temperature, providing behavioral evidence of sensitive thermoreception. Honeybees (*Apis mellifera*) normally choose a temperature range of 35° ± 1.5° C (95° ± 2.7° F); when repeatedly replaced at the warm end of the gradient, individual bees follow their average chosen temperature within ±0.25° C (±0.45° F). Bees also accurately regulate temperature in the hive between 35° and 36° C (95° and 97° F) by behavioral patterns (*e.g.*, beating wings to circulate air) in the brood season.

Among invertebrates other than arthropods, the leech (*Hirudo medicinalis*) can make temperature discriminations with an accuracy of 1° C (about 1.8° F). The slug (*Agriolimax reticulatus*) reacts at temperatures below 21° C (70° F) by increased locomotor activity in response to 0.3° C (0.5° F) cooling over a period of five minutes.

The temperature sensitivity of bloodsucking arthropods (*e.g.*, lice) is considerably greater than that of nearly all other arthropods; the warmth of the victim's body is the primary influence in stimulating and guiding such blood feeders. The so-called castorbean tick (*Ixodes ricinus*), which sucks blood from sheep, responds when its front legs, which are the primary site of thermal sensitivity, are warmed up by 0.5° C (0.9° F). The bloodsucking assassin bug (*Rhodnius prolixus*) responds with direct movement toward any warm stimuli; *e.g.*, when a glass tube warmed 15° C (27° F) above air temperature is kept within about four centimetres (1.5 inches) of its antennae. Similarly, mosquitoes (*Aedes aegypti*) fly readily to a warm, odourless, inanimate surface as if it were that of a warm-blooded creature. The mosquito's antennae are probably the site of the thermosensors, and the animals manifest sensitivity to changes in air temperature of about 0.5° C. In most insects the thermoreceptors appear to be located in the antennae, since they show impairment of thermoreceptive behaviour when part or all of the antennae are removed. Behavioral studies represent a rather gross method of localizing thermosensitive structures, however. A more direct approach to thermoreceptor function in insects has been achieved by electrophysiological methods. Microelectrodes with tips of very small diameter are inserted near the presumed thermosensitive cells. Any electrical nerve impulses elicited by temperature stimuli are amplified and recorded. This method permits, for example, the study and identification in various insects of receptors that are sensitive to cooling. Cockroaches (*Periplaneta americana*) have two whiplike antennae consisting of 120 to 180 ring-shaped segments that grow thinner and longer with increasing distance from the animal's head. There are about 20 cold receptors per antenna; these are located on the thicker segments (Figure 5), with rarely more than one per segment. Each cold receptor consists of a delicate hairlike structure (sensillum) emerging from a ring-shaped wall. The cold sensilla are mechanically protected by large bristles covering the segments of the antenna. At constant temperatures the cold receptor is continuously active, the average maximum frequency of its discharge being about 16 impulses per second at a temperature near 28° C (82° F). At higher and lower temperatures, the steady frequency becomes lower. When the receptor is rapidly cooled, its discharge frequency rises steeply up to 300 impulses per second and then declines gradually to a much lower constant level. On rapid warming, the opposite response is seen; *i.e.*, there is a transient inhibition of the receptor discharge, followed by a gradual restoration of the steady activity. The cold receptor is thus sensitive to constant temperatures as well as to the rate of temperature changes.

Studies with microelectrodes

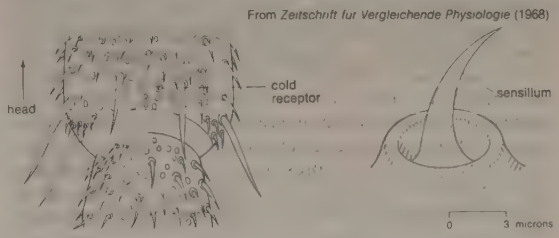


Figure 5: (Left) Portion of cockroach antenna showing location of cold receptor. (Right) Magnification of a single cold receptor of a cockroach.

Caterpillars of various moths (*Lasiocampidae, Saturnidae, Sphingidae*) have cold-receptor cells in their antennae and mouthparts (maxillary palps). Electrophysiological investigations with microelectrodes suggest that just three receptor cells located in the third antennal segment and probably not more than one receptor cell in the maxillary palp are sensitive to cooling. At constant room temperatures, static neural activity from such cells is observed; this activity increases in frequency when the temperature is lowered. During rapid cooling, the frequency rises steeply to a transient maximum of up to 300 impulses per second, while rapid warming produces a temporary inhibition of the discharge. Since only a few cells out of the 20 or 30 that comprise the thermoreceptor structure exhibit the typical electrical response to cooling, specific thermoreceptive function among caterpillars is strongly indicated.

Electrophysiological evidence for the presence of thermosensitive structures also is available for the antennae of honeybees and of migratory locusts (*Locusta migratoria migratorioides*). Temperature-induced changes in the spontaneous electrical activity within the central nervous system of honeybees also have been recorded. Other sen-

sory structures in these animals also can be influenced by temperature, but their primary functions appear to be chemoreceptive or mechanoreceptive.

THERMORECEPTORS IN VERTEBRATES

Fish. Many species of modern bony fish (teleosts) are sensitive to very small changes of temperature of the water in which they live. Various marine teleosts, such as the cod (*Gadus gadus*), have been trained to swim half out of water up a long sloping trough in response to changes of as little as 0.03° to 0.07° C (0.05° to 0.13° F) in the temperature of the water flowing over them.

More detailed conditioning experiments with freshwater fish show that they can distinguish warm from cold, discrimination being made on the basis of thermal change rather than on absolute temperature. Temperature sensitivity persists in these animals when the nerve supplying the lateral line (see *Mechanoreception* above) is cut but is abolished after transection of the spinal cord. When freshwater fish are trained to seek food in response to a change in water temperature, they are found to discriminate differences of less than 0.1° C (0.2° F). Goldfish (*Carassius*) have been trained to discriminate between warm and cold metal rods that have been placed in their tanks. Consistent responses are obtained only when the rod is at least 2° C (3.5° F) colder or warmer than the water. Practically the whole surface of the fish, including the fins, is found to be thermosensitive.

Temperature discrimination in goldfish

This mode of temperature discrimination need not be ascribed to the function of specific thermoreceptors; it could depend on skin receptors that are sensitive to combined mechanical and thermal stimulation. Indeed, electrophysiological recordings from nerve fibres originating in the skin of fish support the latter view. Changes in the electrical activity of these fibres are elicited only when the skin is touched by some solid object; yet the frequency of this mechanically elicited neural discharge is heavily influenced by the temperature of the object used in touching the fish.

From *Journal of Physiology* (1956)

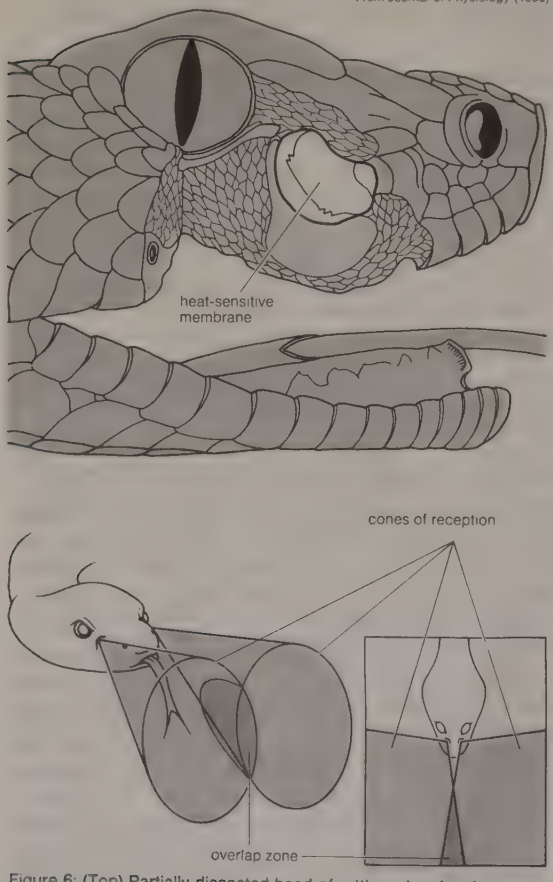


Figure 6: (Top) Partially dissected head of rattlesnake showing heat-sensitive membrane inside pit organ. (Bottom) Cones of reception, directions from which heat energy can be detected (see text).

Elasmobranchs, such as rays and sharks, have distinctive sense organs, called ampullae of Lorenzini, that are highly sensitive to cooling. These organs consist of small capsules within the animal's head that have canals ending at the skin surface. The capsules and their canals are filled with a jellylike substance, sensory-receptor cells being situated within each capsule. Recordings of impulses from single nerve fibres supplying the ampullae of Lorenzini in rays (*Raja*) and dogfish (*Scyliorhinus*) reveal steady activity of the receptors at constant temperatures between 0° and 30° C (32° and 86° F), the average frequency maximum appearing near 19° C (66° F). Rapid cooling causes transient overshooting of the stabilized discharge frequency, while rapid warming produces transient inhibition of the impulses. In some single fibres, cooling by 3° C (5.5° F) leads the frequency to overshoot by about 100 impulses per second. It remains an open question, however, whether the ampullae of Lorenzini are to be called specific thermoreceptors, since they also respond to mechanical stimuli and to weak electrical currents.

Amphibians and reptiles. Rattlesnakes (*Crotalus*) and related species of pit vipers (Viperidae) have a pair of facial pits (Figure 6), sense organs on the head below and in front of the eyes, which are most sensitive thermoreceptors indeed. The pit organs act as directional distance receptors and make it possible for the reptile to strike at warm prey even when the snake's eyes and nose are covered and its tongue has been cut off. Each pit is a cavity about five millimetres deep, equally as wide at the bottom, and narrowing toward the opening at the surface of the head. Inside and separated from the bottom by a narrow air space is a densely innervated membrane of about 10 microns thickness stretching between the walls of the pit. A direct connection between the air space beneath the membrane and the open air maintains equal pressure on both sides of the membrane. Warm-sensitive receptors distributed over the membrane consist of treelike structures of uninsulated (unmyelinated) nerve fibres. Infrared radiation (heat energy) reaches the membrane from an external source through the narrow opening of the pit, permitting the snake not only to detect heat but also to localize coarsely the position of the stimulus. The fields of direction (cones of reception) from which each pit can receive infrared radiation from the environment extend to the front and sides of the head, with a narrow zone of overlap in the middle, as shown in Figure 6.

Under resting conditions, there is an irregular, steady discharge of nerve impulses from the pit organ. Rapid warming by as little as 0.002° C (0.004° F) at the nerve endings elicits a significant increase in impulse frequency; cooling produces an inhibition of the resting discharge. In contrast to the warmth receptors in mammals, the reptile's pit receptors are practically insensitive to steady temperatures, despite their high sensitivity to rate of thermal change. The distinctive consequence in the snake's adaptive behaviour is that gradual variations in air temperature tend to occur without detection, only the more rapid changes in infrared radiation being discriminated. Sensitivity to rapid temperature changes is enhanced by the very limited heat capacity of the receptive membrane (since it is so thin). When an animal that is 10° C (18° F) warmer than the environmental background appears for half a second at a distance of 40 centimetres (about 16 inches) in front of the snake, the heat energy radiated is enough to elevate significantly the frequency of receptor discharge in the pit organ. Indeed, behavioral experiments show that under these conditions the snake is able to discover warm prey through the victim's infrared radiations.

Detection of warmth of their prey by snakes

As poikilotherms, reptiles have practically no internal neural or metabolic mechanisms for maintaining their body temperature within physiologically safe limits. Nevertheless, such reptiles as snakes and lizards are able to keep their body temperature near these safe levels through behavioral regulation (*i.e.*, by moving to cooler or warmer places as necessary). The body temperatures of two samples of lizards (*Sceloporus magister* and *Cnemidophorus tesselatus*), for example, were found to be 34.9° ±0.6°

C (94.8° ± 1.1° F) and 41.3° ± 0.2° C (106.3° ± 0.4° F), respectively, although the average air temperatures were 33° C (91° F). Highly accurate regulation is recorded for a snake (*Crotalus cerastes*) that moved partially in and out of its burrow into the sun to maintain a body temperature of 31° to 32° C (88° to 90° F) over several hours. The desert iguana (*Dipsosaurus dorsalis*) regulates its body temperature largely by behavioral mechanisms to achieve and hold body temperatures near 38.5° C (101.3° F). These adjustments by iguanas include postural orientation to solar radiation both inside and outside burrows and altered thermal contact of the body surface with the soil. Although supporting direct evidence remains to be more fully developed, it appears that reptiles have thermosensitive nerve structures in the brain as well as in the skin.

There is some electrophysiological evidence of thermal sensitivity among amphibians, but only to relatively large temperature changes. The lateral-line organs in frogs (*Xenopus laevis*), which, as in fish, are sensitive to minute water turbulence, also respond to static temperatures and to temperature changes. Whether these responses have any adaptive, behavioral significance for temperature detection remains to be demonstrated. It has been reported, however, that a frog placed in a pan of cool water will not jump out as the pan is heated, if the temperature changes are gradual enough. Indeed, frogs are recorded to remain in the water this way until they are boiled to death.

Birds. Birds are homoiothermic, normally maintaining their body temperature within a range of less than 1° C (1.8° F). Investigations of temperature regulation in birds suggest the existence of thermosensors both in the lower part of the brain (hypothalamus) and in the skin.

Direct electrophysiological evidence of thermoreceptors has been obtained in the tongues of chickens and in the skin of pigeons by recording from individual fibres of nerves serving the receptors. At a constant temperature of 20° C (68° F), a high level of static activity was observed for cold receptors in the chicken's tongue. When the temperature of the tongue was maintained at 44° C (111° F), individual cold fibres showed a low, steadystate frequency of two to four impulses per second. A temperature drop of 9° C (16° F) was found to elicit an initial response of 30 impulses per second, which gradually declined to a new static frequency of eight impulses per second. Rewarming of the tongue resulted in a cessation of detectable electrical activity for several seconds; no specific warm receptors were found. There is some electrophysiological evidence of cold and warm receptors in the skin of pigeons.

Megapodes, large-footed birds such as the Australian mallee fowl (*Leipoa*), or brush turkey, bury their eggs. They depend on the thermal sensitivity of their face or mouthparts to guide their efforts in controlling the temperature of the eggs during hatching. The eggs are incubated in mounds where heat is generated through the fermentation of rotting vegetation and by irradiation from the sun. For extended periods of time the male bird is busy covering and uncovering the eggs, normally keeping the temperature almost constant at 34° ± 1° C (93° ± 2° F) over the unusually long incubation period (as much as 63 days) that characterizes this family of birds.

Mound building to regulate nest temperatures

Mammals. Detailed information is available from electrophysiological investigations of single thermosensitive nerve fibres in the skin of mammals, particularly cats and monkeys. The nose of a cat contains numerous cold and warm receptors that are highly specific in responding to thermal stimuli; they are not excited by mechanical deformation of the skin. As a rule, each thermoreceptor is connected with a single nerve fibre. By using very finely tipped thermal stimulators, investigators can locate precisely the sites of warm and cold receptors in the skin; the details of the underlying cellular structure at these spots have been studied by electron microscopy. At the site of a cold-sensitive spot in the cat's nose, a thin, myelinated (insulated) nerve fibre penetrates the dermis and divides into several unmyelinated branches about 70 microns beneath the skin surface (Figure 7). The tips of these branches have been shown to be the cold-sensitive nerve endings proper; they come into close contact with the basal cells of the epidermis. In most cases the nerve endings are embedded

Figure 7: Cold receptor in the skin of a cat as seen with electron microscope.

in small concavities on the lower surface of the basal cells. Warm receptors remain to be identified; they appear to be situated in a deeper layer of the skin. In monkeys and, presumably, in man, warm receptors are innervated by unmyelinated nerve fibres (diameter one micron, impulse conduction velocity of 0.5 to 1.5 metres [1.5 to five feet] per second); cold fibres are served either by unmyelinated fibres or by thin myelinated fibres (diameter two to four microns, conduction velocity three to 20 metres [ten to 66 feet] per second).

At constant skin temperatures in the normal range, cold fibres in mammals are found to be continuously active. The average maximum frequency of the static discharge is observed at 27° C (81° F), the extreme limits of the normal range of static activity being 5° and 42° C (41° and 108° F). At skin temperatures above 45° C (113° F), cold receptors again can be activated. This so-called paradoxical discharge corresponds to the similar paradoxical sensation of cold in man when a hot object is touched or the hand is put into hot water. On sudden cooling, temporary overshooting can be 30 times higher than the static frequency of eight impulses per second. Warm receptors in the cat's nose start to show their activity at skin temperatures of 30° C (86° F) and reach an average maximum of 35 impulses per second at 46° C (115° F). Above this temperature the discharge suddenly falls off.

Similar populations of cold and warm receptors have been found in monkey skin, which has an additional group of warm fibres distinguished by an average static maximum at 41° C (106° F). The transient overshooting of warm-receptor activity can be several times higher than the static maximum frequency. When cold stimuli of the same magnitude (*e.g.*, 3° C [37° F]) are rapidly applied to the warmer skin, the degree of overshooting reaches a maximum at a skin temperature of 27° C (81° F). This temperature corresponds to that which elicits maximum static discharge from the cold fibres. Overshooting on rapid warming of warm receptors follows the same general rule: the temporary maximum occurs at temperatures for which the static discharge frequency also appears at a maximum.

Distinctive properties of cold receptors are found in hibernators. European hamsters (*Cricetus cricetus*) tend to maintain body temperatures of about 5° C (41° F) during hibernation. Further cooling, however, elicits arousal reactions from the animal that indicate thermoreceptive function is intact. Electrophysiological investigations have shown that myelinated cold fibres serving the hamster's

nose are continuously active at very low temperatures, having a static maximum near 4° C (39° F). In contrast to these findings, myelinated cold fibres in mammals that do not hibernate are blocked at temperatures below 10° C (50° F).

Relatively little is known about the processing of information from skin thermoreceptors in the central nervous system (brain and spinal cord). Responses to cooling the tongue have been recorded from single nerve cells (neurons) of the brain's thalamus in monkeys. In addition to a few brain neurons that are excited both by mechanical stimulation and cooling of the tongue, there are also numerous nerve cells in the thalamus that respond only to cooling. The latter neurons exhibit a static discharge in the temperature range (at the tongue) between 15° and 44° C (59° and 111° F), the maximum frequency being at 21° to 31° C (70° to 88° F). Rapid cooling of the tongue causes considerable transient overshooting in frequency from thalamic nerve cells over the entire range of preadapting temperatures used; by contrast, warming of the tongue results in a transient inhibition but not an increase in the rate of a discharge of any nerve elements in the brain. Thus, the activity of thalamic nerve cells activated by cooling the tongue closely reflects the behaviour of the peripheral (*e.g.,* tongue or skin) cold receptors. Even in the outer layers of the brain (cerebral cortex), nerve cells that respond specifically to cooling of the skin have been found. Other individual cortical neurons in the cat, however, receive signals not only from peripheral thermoreceptors but from mechanoreceptors and taste receptors as well.

Thermal information processing in the monkey brain

Cats can be trained to respond behaviorally to thermal stimulation (*e.g.,* they will press a bar or lever). These experiments reveal cats to be relatively insensitive in discriminating warm and cold stimuli applied to the furred skin of the trunk or the legs, requiring temperature differences amounting to several degrees Celsius. This does not necessarily mean that cats have no finer thermal sensitivity than their grossly observable behaviour suggests. Indeed, autonomic regulatory responses, such as increased blood flow to the ear, can be elicited by mild warming of the paws, even though such warming is inadequate as a signal for training behavioral responses. In contrast to the low thermal sensitivity of their skin elsewhere in the body, cats have been found to manifest behavioral responses when the nose and upper lip are warmed or cooled by only 0.1° to 0.2° C (0.2° to 0.4° F). This corresponds to levels of thermal sensitivity of the face in human subjects and also is in accordance with electrophysiological evidence of high thermal sensitivity of the cat's nasal region.

Mammalian thermoreceptor structures, each containing elements sensitive to warm and cold, are to be found in the skin, in the deep tissues of the body, and in the hypothalamus and spinal cord. While much of the evidence of thermosensors in the central nervous system derives from experiments in neural temperature regulation within the body, in 1963 microelectrodes were used to record directly the activity of thermosensitive neurons in the frontal part of the cat hypothalamus. Many investigations made with this method show that most neural elements in the cat and rabbit hypothalamus are practically insensitive to directly applied temperature stimuli. There are, however, two smaller populations of thermosensitive neurons, one of which responds to local warming and the other to local cooling of the brain tissue.

Warm-sensitive brain neurons in cats and rabbits increase the frequency of their static impulse discharge in direct proportion to the degree that hypothalamic temperature is raised above normal. Similarly, the cold receptors there respond with an increase in impulse frequency when the temperature of the hypothalamus falls below the normal value for the animal's body. Since temperature in the deeper tissues of the body (*e.g.,* the brain) varies quite slowly, the activity of hypothalamic thermosensors seems to be almost entirely a function of the level of temperature alone and not of the rate of temperature change. By contrast, the thermoreceptors in mammalian skin are highly sensitive to rapid changes in temperature. Intervening elements in the nervous system have been identified that integrate temperature signals from the hypothalamus and from the skin. Thermosensors are also found to be localized in the midbrain of rabbits and in various parts of the spinal cord in guinea pigs and dogs. These findings are in good agreement with observations that thermoregulatory responses such as shivering or panting can be influenced by local temperature changes in the spinal region of the animal.

Signals from hypothalamic thermosensors, from deep-body thermosensors, and from those in the skin are integrated in thermoregulatory centres located mainly in the mammal and bird hypothalamus. The integrated signals provide information about the inner (core) temperature of the body and about thermal changes at the periphery (body surface). Such information serves to activate internal mechanisms that maintain body temperature within the normal range of values. When signals from warm receptors (especially from those in the hypothalamus) prevail over signals from cold sensors, such heat-loss mechanisms as sweating, panting, and widening of blood vessels (vasodilatation) in the skin act to reduce body temperature. When signals predominate from cold receptors (particularly from those in the skin), heat-conservation mechanisms are initiated. Heat production rises as muscles expend energy in shivering and through other metabolic reactions (nonshivering thermogenesis); heat loss is reduced by mechanisms that narrow the blood vessels (vasoconstriction) in the skin and that fluff out hairs or feathers to enhance thermal insulation. All these involuntary, or autonomic, regulatory changes continue even without the function of the cerebral cortex; thus, they do not require consciousness, persisting during light anesthesia or during sleep.

Sweating, panting, and shivering

In human beings, the function of thermosensors is closely involved in the highly emotional experiences of thermal comfort and discomfort. Whereas temperature sensations are mainly related to the activity of warm and cold receptors in the human skin, thermal comfort and discomfort reflect the general state of the thermoregulatory system, involving signals not only from thermoreceptors in the skin but from thermoreceptors in deep-body regions and in the hypothalamus as well. Thus, the same temperature at the skin can be experienced as comfortable or uncomfortable, depending on the thermal condition of the person's whole body. When one is overheated, an ice bag applied to the head may be perceived as pleasant; but, if someone is generally chilled just to the point of shivering, the same cold stimulus can be most unpleasant. (H.He.)

Chemoreception

All animals react to chemicals in the environment, initially through a sensory process called chemoreception. The process begins when chemical stimuli come in contact with chemoreceptors, specialized cells in the body that convert (transduce) the immediate effects of such substances directly or indirectly into nerve impulses. A nerve cell (neuron) that makes a direct conversion is called a primary receptor; a cell that is not a neuron but that responds to stimulation by inducing activity in an adjacent nerve cell is called a secondary receptor.

CLASSES OF CHEMORECEPTORS

In man two distinct classes of chemoreceptors are recognized: taste (gustatory) receptors, as found in taste buds on the tongue; and smell (olfactory) receptors, embedded high in the lining (epithelium) of the nasal cavity. These respond to different classes of chemicals: gustatory receptors to water-soluble materials (*e.g.,* salt) in direct contact with them and olfactory receptors to generally water-insoluble, vaporous materials that may arise from a distant source, such as a neighbour's kitchen. The receptors themselves are also different; gustatory receptors are specialized epithelial cells (secondary receptors) with neurons branching among them, while olfactory receptors are nerve cells (primary receptors) with fibres leading to the brain.

In all air-breathing vertebrates (*e.g.,* reptiles, birds, and mammals) the two classes of chemoreceptors are easily identifiable. In fish gustatory organs are on the fins and even the tail, as well as in and near the mouth, all still

Fish that taste with their tails

recognizable as taste buds. The nostrils in fish do not usually open into the mouth, but they are lined with olfactory epithelium. Much lower concentrations of chemicals are needed to elicit responses in fish for smell than for taste. These concentrations are similar to those for air breathers, permitting separate identification of the chemical senses for aquatic and terrestrial vertebrates.

For some invertebrates (*e.g.*, worms), however, distinctions between taste and smell receptors may not emerge. Chemoreceptors of these animals are structurally different from those of vertebrates, and their locations on the body are different. It has been held that invertebrate animals have only one chemical sense, with different sensitivities for various chemicals, as measured by the lowest concentrations (thresholds) of chemicals that can be received. Terrestrial invertebrates, particularly insects, do exhibit separable chemoreceptive capacities, however; additional study seems likely to reveal similar distinctions for other invertebrates. For these animals, the terms distance chemoreceptors and contact chemoreceptors are preferred by many biologists over the terms (*e.g.*, smell and taste) used in human physiology. Separation of these seems feasible because contact chemoreceptors are usually stimulated by nonvolatile, water-soluble chemicals, while distance chemoreceptors typically respond to volatile, oil-soluble chemicals. In addition, thresholds for stimulation of distance chemoreceptors are usually very much lower than those for contact chemoreceptors. Generally the behavioral results of contact chemoreception are feeding, mating, or the deposit of eggs, while those of distance chemoreception are orientation or movement of the animal toward or away from a volatile chemical.

Aquatic animals and terrestrial species with mucus-secreting skins are generally sensitive to chemicals all over the body, reacting with avoidance. This sensitivity has been called the common chemical sense. Man and other terrestrial vertebrates have a remnant of this receptor system that responds to irritants in the mucous membranes of the mouth, eyes, and genital organs. Common chemical receptors are thought to be free nerve endings (branching structures, or dendrites, of nerve cells) in the skin or in moist membranes. Even on the basis of relatively few studies, the common chemical sense is known to be separable from the sense of pain, and thus it is considered as a separate sensory capacity.

Receptors for humidity, particularly well studied in insects, may or may not be chemoreceptors. There is no question that some animals can orient toward or away from regions of high or low atmospheric humidity. The question is whether this is true hygroreception (*i.e.*, stimulation of the receptor by moisture-saturation deficit) or is stimulation by water acting as an odorous chemical. While the matter is far from settled, it seems that some insects and possibly mammals actually may be able to smell water, while others have true hygroreceptors.

Flavour

In common speech the word taste refers to what is more correctly designated as flavour. For man, flavour sensations represent integration by the central nervous system (*e.g.*, the brain) of a complex of stimuli: gustatory, olfactory, common chemical, tactile, thermal, even painful. When carefully studied in other species (*e.g.*, a few other mammals and a few insects), reactions to foods seem to be similar to those of man, with multidimensional stimulation involved in food preferences.

ADAPTIVE FUNCTIONS OF CHEMORECEPTION

For most animals, chemical stimuli are leading sources of information about the environment; even man relies heavily on chemoreception for food selection. Species identification, mate finding, courtship, and mating are also chemically directed among most animals.

Food procurement. Foods are generally located by reception of odours they emit, sampled for palatability by both contact and distance chemoreception, fed upon only if they supply appropriate chemical stimuli during feeding, and laid aside either when the animal is full or when the animal's threshold of response for the stimulating chemical rises above the intensity of stimulation provided by the foods.

At least four classes of chemicals are recognized that affect feeding behaviour: (1) attractants: odours eliciting movement *toward* the source; (2) repellents: odours that prompt the animal to move *away* from the source; (3) feeding stimulants (phagostimulants): tastes and odours that induce the animal to feed; and (4) feeding deterrents (antifeedants): tastes and odours that inhibit feeding behaviour. Chemicals in foods that attract animals or that induce feeding are not necessarily nutritionally valuable in themselves; in food plants, the stimulants often are so-called secondary plant substances (*e.g.*, odorous essential oils) that provide little nourishment. Among animals that are preyed upon as food, the stimulants are often traces of odorous materials present on the body surface. Indeed, animals will feed on nutritionally worthless materials that have been experimentally impregnated with appropriate phagostimulants. Ordinarily, however, specific feeding stimulants are part of an animal's natural food (see also BEHAVIOUR, ANIMAL: *Feeding Behaviour*).

Symbiotic relationships. Most parasites do not just blunder onto their hosts but, rather, orient themselves toward suitable animals or plants. Little is known about the guiding stimuli for most parasites, but for some the odour of the host acts as an attractant, and the taste of the host's body surface functions as a feeding stimulant.

Chemoreception among parasitic wasps

Parasitic wasps that lay their eggs on wood-boring insects, for example, locate their targets in logs through olfactory signals. The wasp then drills into the log with a complex egg-laying structure (ovipositor) on the end of which are contact chemoreceptors that allow the insect to sample the prospective host to determine whether or not it is already parasitized. Animals that establish nonparasitic (mutualistic or commensalistic) relationships also find each other by chemical clues; or at least the mobile member of a pair finds the nonmobile member through chemoreception. Sea anemones that attach themselves to shells housing hermit crabs, for example, detect the proper shells with contact chemoreceptors on their tentacles. Annelid worms that are commensal (feeding together) with starfish or sea urchins to which they cling locate the latter by chemicals given off by the hosts.

Communication. Many animals release chemicals that influence other individuals behaviorally or at least physiologically. Usually produced by glands, these chemical communication signals have been named pheromones because they seem to act somewhat like hormones inside an animal's body. Females of some moths, for example, produce scents that attract males from great distances (a behavioral effect).

Ovarian suppression in bees

Queen honeybees give off a chemical (so-called queen substance) that suppresses ovarian development in worker bees (a physiological effect). Basically the general classes of information that are coded in chemical signals are concerned with species or individual identification, with social communication, and with sexual or reproductive activity.

In aggregating as groups or in dispersal, animals depend on their ability to identify species or individuals. Thus, honeybees scent-mark their own hive and areas around it with odours that uniquely identify that particular insect community for its members. Many mammals are individually territorial, marking the boundaries of their territories with special glandular secretions (*e.g.*, deer), with body odours (*e.g.*, bears), or with urine (*e.g.*, dogs).

Chemical signals facilitate cooperation among social insects and many mammals. When their colony is endangered, for instance, ants, bees, and wasps alert the group with alarm odours. They also deposit chemicals that serve as guidance signals to indicate the way to sources of food or to living quarters.

Most of the sexual signals that animals produce at all stages of mating are chemical. Females of many mammalian species, for example, produce specific odours that attract only males of the same species. Male bumblebees mark leaves or sticks with a scent that induces females of their species to tarry for mating. In many species mating itself is stimulated in one or both sexes by special chemicals produced by the partners. Male tree crickets, for instance, produce a glandular secretion on which the female feeds during mating.

Orientation. Besides being oriented toward or away from food or mates, many animals are guided to suitable habitats by chemicals emanating from plants or from other environmental features. Fish such as salmon, which return from the ocean to lay their eggs in fresh water, generally come back to the specific stream where they themselves were hatched, guided by the odour of the stream. Other fish recognize their nesting areas by odours produced by plants in the vicinity.

Protection against predators. A most effective form of chemical protection is found in marine slugs and snails that produce strong acid secretions when disturbed. These secretions can injure other animals. Many species of animals produce chemicals that are repellent without necessarily being dangerous; for example, stinkbugs, millipedes, skunks, and some earthworms produce strongly smelling or bitter-tasting secretions when disturbed. An animal that causes a predator to become ill long after contact is not thereby directly protected. If the prey has a special taste or smell, however, the predator that samples it and later sickens learns to avoid the taste or smell, thus sparing other members of the species upon which it might otherwise prey.

Repellents of stinkbugs and skunks

CHEMORECEPTORS IN LOWER INVERTEBRATES

Detailed evidence of chemoreception is available for only insects and mammals. Indeed, chemoreception has been studied in depth for only three or four species of insects and four or five species of mammals. For most animals data for secure generalizations are lacking.

Protozoa. Protozoans, even though they are single-celled, behave as if they had a nervous system. They are sensitive to chemicals in the environment and usually select some foods in preference to others. Carbon dioxide dissolved at low concentrations attracts many protozoans and may be the agent that leads them to foods. Some protozoans (*e.g., Spathidium*), however, can locate specific foods at a distance, presumably by a chemical sense. Ciliates (*e.g., Paramecium*) are most sensitive to chemical stimulation at the anterior (front) end; the receptors are probably special cilia (hairlike structures). *Paramecium* takes nonfoods, such as carmine particles, but soon "learns" to stop this, the change in behaviour persisting for some days. In some ciliates (*e.g., Vorticella*) that reproduce by exchanging genetic material between individuals (conjugation), a motile partner (conjugant) swims to a stationary individual. The swimmer is attracted from up to a millimetre away by a chemical produced by the fixed partner. All of these behaviour patterns performed by only one cell are nevertheless similar to those of multicellular animals.

Cnidaria. Chemoreception is doubtless the most crucial receptive capacity of cnidaria (*e.g., Hydra* and jellyfish), but little is known about the organs involved. Sensitivity to food chemicals is greatest near the mouth and tentacles, but specialized organs remain to be described. Almost all receptors are free nerve endings in the integument (body surface). *Hydra* exhibits feeding behaviour when stimulated by such chemicals as reduced glutathione or tyrosine. This reaction occurs in about half of the tests with weak solutions (1×10^{-6} molar) of these substances. Reduced glutathione acts similarly on the Portuguese man-of-war (*Physalia*) and some other coelenterates called marine hydroids. Amino acids other than tyrosine induce a feeding response in some coelenterates: valine and glutamine in sea anemones and proline in some hydroids and corals.

The feeding sequence of coelenterates is highly coordinated, despite the presence of only a very primitive kind of nervous system called a nerve net. Contact with food causes discharge of stinging or entangling structures (nematocysts), the reaction being released by a combination of chemical and tactile stimuli. The tentacles then draw the prey into the mouth. This response may be evoked by release of glutathione or amino acids from the injured prey.

Other behaviour patterns of coelenterates have been little studied. Anemone fish (*e.g., Amphiprion*) live safely among the tentacles of sea anemones that kill other fishes. Seemingly the mucous coat of the anemone fish develops a chemical that inhibits the discharge of nematocysts, although other interpretations of observations made so far are possible. Many marine coelenterates that live in immobile groups shed sperms or eggs (depending on their sex) synchronously, the activity probably being regulated by chemicals given off by some individuals that trigger discharge in others. A swimming sea anemone, when touched by a starfish that feeds upon it, releases its hold and swims away. Identification of the predator starfish is specifically chemical. Reactions of coelenterates to chemical stimuli are far from stereotyped, a wide range of responses being observable.

Platyhelminthes. Flatworms (Platyhelminthes) have two major life-styles—free-living (turbellarians) and parasitic (tapeworms and flukes)—and their reactions vary accordingly.

For some free-living flatworms (*e.g.,* freshwater planarians) the locations of chemoreceptors in the body are known, but their structure is not. Planarians locate foods at a distance, and their behaviour during this process indicates that earlike protuberances (the auricles) on the head bear the receptors. Water currents elicit orientation movements, the animals crawling upstream when thus stimulated, as if they were making an olfactory response. Removal of a structure called the auricular groove abolishes planarian responses to foods; the receptor organs in the groove are thought to be ciliated glandular patches of nerve cells. Upon reaching food, the worm makes contact with its anterior end and with the tip of its pharynx (proboscis). Ingestion then may or may not occur, the reaction resembling selective taste (gustatory) responses of other animals. The tip of the worm's proboscis has receptors; indeed, an isolated pharynx cut away from the rest of the body will feed on appropriate foods.

Reactions of flatworms to pure chemicals

Flatworms have been experimentally subjected to stimulation with many pure chemicals, most at concentrations not likely to be encountered in nature. The animals are usually attracted by relatively weak solutions and repelled by high concentrations. They respond to natural food juices and experimentally to pure amino acids and their derivatives. A worm called *Dugesia* reacts positively to such chemicals as lysine and glutamine, negatively to aspartic acid, asparagine, and α-keto-glutaric acid, and gives no observable response to hydroxyproline and glutamic acid. Planarians of different species, when mixed together in the same tank of water, can be separated by species through differences in their chemical-recognition behaviour. These distinctive chemically mediated reactions indicate well-developed sensory function for the planarian nervous system.

Little evidence is available about chemical sensitivity among tapeworms and flukes. Tapeworms are said to have only tactile organs, but supporting evidence is almost nil. Adult flukes obviously find their way to specific organs in the bodies of animals they parasitize, but the sensory mechanisms are unknown. The free-swimming stages (miracidia and cercariae) in the life cycle of flukes find their hosts effectively, but there is no general agreement on how this is done. Some workers hold that they swim at random and enter whatever body they encounter; others say that the flukes swim at random but select the host on contact; still others claim that they orient toward the host before contact. Perhaps different species of flukes vary in their behaviour, but the evidence is too sparse to draw general conclusions.

Nematoda. For a phylum with so many commercially and medically important parasites (as well as free-living species), the lack of studies on chemoreception in roundworms (nematoda) is surprising. The integument of these roundworms is supplied with many types of receptors, mostly free nerve endings. These are concentrated anteriorly, particularly on structures around the mouth called papillae. Nematode papillae could be chemoreceptors, but the possibility is supported by no direct evidence. Some roundworms have specialized glandulo-neural structures (amphids at the anterior end of the body and phasmids at the posterior end) that have been claimed to be chemoreceptive, again without critical verifying evidence.

Except for nematodes that parasitize plants, no agree-

ment has been reached on how these animals find their hosts or foods or how they form "social" aggregations, as some free-living species of roundworms do. Parasitic nematodes may attack the roots of plants in response to a chemical attractant in the roots. In some cases the attractant is found to be carbon dioxide that stimulates the worms at a distance, with some other chemical acting on contact. The possibility that control of some agriculturally destructive pests may be achieved by changing the chemical environment in the soil is drawing increased attention to behavioral studies of these nematodes.

Echinodermata. These marine animals (*e.g.*, starfish, sea urchins, sea cucumbers) have also been little studied. They are generally sensitive to chemicals, seemingly most acutely at the tips of their myriad tubular "feet" (podia). Only free nerve endings are present in the integument (skin) of most echinoderm species, but sea cucumbers have sensory pits on their tentacles with more specialized nerve endings. The concentration of primary sensory cells in the integument of many echinoderms is truly striking, upwards of 4,000 per square millimetre (2,600,000 per square inch) being reported for certain starfish. These endings may be multisensitive (to a number of chemicals), or they may be functionally differentiated although structurally they appear to be identical.

The chemoreceptive "feet" of starfish

Reports of studies of chemical reactions among echinoderms are few and spotty. These animals respond positively to natural foods and to some food chemicals (such as glutamic acid) at a distance, and they feed on specific items on contact. They avoid harmful chemicals (*e.g.*, injurious acids and salts). They also form specific aggregations, possibly through chemical responses to their fellows, and are known to spawn synchronously as a result of chemicals released during the process.

Annelida. Annelids (*e.g.*, leeches and earthworms) are sensitive to chemicals all over the body; they are selective in feeding, but no specialized chemoreceptors are yet known for them. Three types of nerve endings in the skin of these animals have been claimed to be chemoreceptive, but without direct evidence: (1) primary sensory cells concentrated at the anterior end, up to 700 per square millimetre (450,000 per square inch) in front of the mouth (on the prostomium) of an earthworm; (2) branching free nerve endings in the skin, possibly mechanoreceptors rather than chemoreceptors; and (3) special concentrations of nerve endings, called integumental sense organs. Some "hairy" marine annelids (polychaets) have a so-called nuchal organ near the head, ranging in complexity from a simple ciliated pit to an elaborate set of folds covering many of the ringlike segments (somites) that form the body. The nuchal organ has been reported as chemoreceptive, but no direct evidence has been produced.

Chemoreception among annelids has been studied mainly by dipping them into or flooding them with various solutions and noting withdrawal or by feeding them natural and man-made foods. The animals respond appropriately, so that thresholds for eliciting responses have been determined. What these mean in the lives of the worms is generally obscure; as usual, low concentrations of many substances are accepted or produce positive responses, whereas high concentrations are rejected or repel. Studies of nerve impulses picked up from receptors in the skin of the body wall have been made with earthworms. The receptors, still unidentified, produce impulses when stimulated with appropriate concentrations of table salt, quinine, and acids, but they fail to respond to ordinary sugar (sucrose). The prostomium, however, does have receptors that are sensitive to sucrose solutions.

Feeding, selection of places on which to settle by some marine annelids, and selection of soil by earthworms have been shown to be chemically mediated. Commensal polychaetes (*e.g.*, *Podarke*) distinguish the organisms with which they live through chemicals coming from their hosts. Synchronous spawning occurs in many anchored (sessile) marine worms, being mediated through the release of signal chemicals. Release of sperms by breeding males of *Platynereis*, a swimming marine polychaete, requires chemical stimuli from the female. Earthworms incorporate an alarm chemical in the mucus given off when they are roughly handled; the effect is to repel other earthworms for as long as several months thereafter.

Mollusca. More information about chemoreception among mollusks (*e.g.*, snails, clams, squids) is available than there is for the groups discussed so far; but these animals comprise a large phylum, and very few species have been studied.

Chemical sensitivity is generally distributed over the mollusk's body, being greatest at the mouth, tentacles, front of the foot, and along the edge of its thin, capelike mantle. The receptors, although not identified with certainty, are thought to be variously branched free nerve endings. Body regions known to be most sensitive to chemicals have high concentrations of these cells. These regions are: (1) tentacles—a variety of projections on various parts of the body; (2) osphradia—ridges or projections near the front of the mantle cavity, best studied in marine gastropods (*e.g.*, snails and slugs); (3) abdominal receptors at the base of the siphons in bivalves (*e.g.*, oysters and mussels); and (4) olfactory pockets behind the eyes in cephalopods (*e.g.*, octopuses and nautiluses). Other organs have been designated as chemoreceptors, but with no critical evidence: (1) so-called subradular organs in the mouths of lower mollusks; (2) a structure called Hancock's organ in some gastropods; and (3) rhinophores (once identified as "olfactory" tentacles) of some gastropods called opisthobranchs. The last, however, are almost certainly established as receptors for water currents rather than as chemoreceptors.

Most of the physiological studies with mollusks have been on reactions to food or to foreign chemicals. Octopuses have been blinded and then trained or conditioned to respond to pure chemicals with specific behaviour patterns. Studies of orientation to or acceptance of feeding stimulants have shown that tentacles and osphradia bear receptors for odorous materials and that receptors near the mouth initiate feeding. Thus separation of contact from distance chemoreception among these animals seems probable; but, until specific receptors are identified through their nerve impulses, the distinction remains conjectural. Although nerve-impulse studies have been made with at least two gastropods (*Aplysia* and *Buccinum*), specific receptors have not been identified thus far. The osphradium has finally been shown to bear chemoreceptors (a matter long debated), and reactions to food extracts and chemicals in natural foods have been studied.

Blinded octopuses

Location of food or prey by many species of mollusks involves what suggests distance chemoreception, generally through the tentacles. Some carnivorous land snails detect and follow (by "tasting") the slime trail left by the prey. Specific "social" aggregations are common among marine bivalves; some of these are brought about by the settling of bivalve larvae near chemically detected members of their group (conspecifics). Chemically regulated synchronous spawning is common among marine mollusks. Land snails and slugs find mating partners by following their slime trails by "tasting" them. Limpets and other snails that live close to the shore emerge to feed when seawater splashes on them at low tide; the sense organs involved differentiate seawater from rain.

Many bivalves and gastropods react strikingly to chemicals from their predators. Herbivorous marine snails, for example, move rapidly away from predators as soon as they touch them. A freshwater snail (*Physa*), when touched by a leech, swings its shell back and forth and then drops to the bottom. These reactions are induced by specific chemicals; the skin of echinoderms, for instance, has yielded such a material, the extract being found to resemble a group of chemicals called saponins.

ARTHROPOD CHEMORECEPTORS

In the Arthropoda, which includes more than two-thirds the total number of all individual animals alive, detailed chemoreceptive studies have been reported for less than 10 species of insects and five species of crustaceans; reliable information about other arthropods (*e.g.*, sow bugs and centipedes) is rudimentary. Many of these latter animals have hairs on their outer surface (exoskeleton) that may be chemosensory, since they are similar to those known to be chemoreceptive in insects and crustaceans.

Responses to food and mates, supposedly chemically mediated, have been described for millipedes, centipedes, and a number of arachnids (*e.g.*, spiders). Electrophysiological studies of chemoreceptors have been made with the horseshoe crab (*Limulus*) found on many beaches. The receptors are in spines on the legs and chilaria (flaps behind the mouth) of the animal. Each sense organ has from six to 15 nerve cells that respond or fire when bathed in clam juice or in solutions of amino acids. A tick (*Ornithodoros*), when fed through an artificial membrane, accepts glucose solutions with such substances as reduced glutathione, adenosine triphosphate, and nicotinamide-adenine-dinucleotide; glutamic acid inhibits feeding behaviour in this arachnid. Among some wandering spiders, the male locates the female by the scent of her silken dragline, which serves to identify species and sex. Contact chemoreceptors at the tips of the spider's legs are the sensitive structures. These observations represent a good sample of the scattered work to date with arthropods other than insects and crustaceans.

Chemoreceptors of horseshoe crabs, ticks, and spiders

Crustacea. Crustaceans include such arthropods as crabs, lobsters, shrimps, barnacles, and many other forms. For a number of crustacean species, reactions to food chemicals or other substances have been used to locate the body regions that bear chemoreceptors. The list is impressive. Distance chemoreceptors are borne on the antennae and the smaller antennules, specialized structures (esthetascs) on the tips of the antennules being particularly sensitive (Figure 8). Contact chemoreceptors are borne chiefly on the tips of the walking legs, the mouthparts,

Hubert Frings

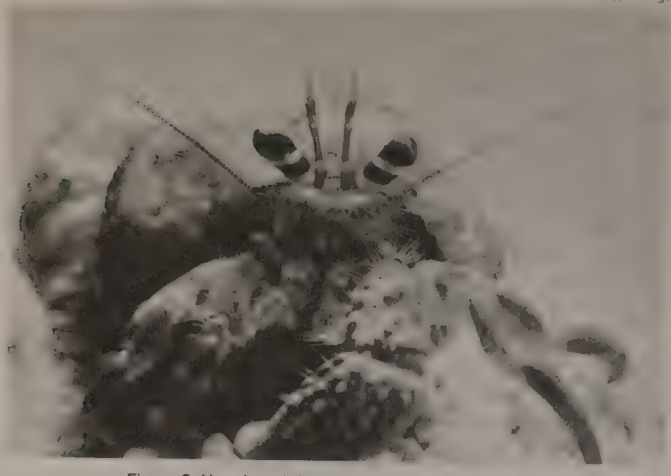


Figure 8: Hermit crab in shell, showing antennae (long and thin) and antennules (held vertically between eyes) with esthetascs (specific chemoreceptors) along edges near tips.

antennules, tail flap (telson), walls of the gill chambers, and, in some species, on the general body surface.

Locations and structure of chemoreceptors. The sense organs in these regions are various, but only the esthetascs have been shown electrophysiologically to be chemoreceptive. Scattered over the body are so-called funnel canals (or pore organs), which are assumed to mediate avoidance reactions to high concentrations of chemicals. Also widely distributed over the body is a variety of hairlike structures that are similar in appearance to known chemoreceptors of insects. Short blunt projections, resembling certain specialized receptors (basiconic sensilla) of insects, on the body wall of terrestrial isopods (*e.g.*, wood louse or pill bug) are also assumed to be chemoreceptive. The esthetascs at the tips of the antennules are groups of hairlike or spinelike structures. Receptors in these produce nerve impulses when stimulated with a variety of chemicals. Each esthetasc hair receives 100–500 nerve endings from cells aggregated in a ganglion-like structure at its base. The nerve endings have a cilia-like pattern of fibrils, characteristic of the primary chemoreceptors of insects and vertebrates. The outer layer (cuticle) of the esthetascs is very thin, but it has no openings through it, as does the cuticle of the sensory hairs of insects.

Most studies on chemoreception among crustaceans have been made on a few species of crabs and crayfish, with food selection or reactions to chemicals as indicators of reception. Tests before and after removal of parts of the body have led to the discovery of the chemoreceptor locations. There have been a few recent electrophysiological studies with only a very limited number of species.

Responses. In general, crustaceans respond to a wide range of chemicals, negatively at high concentrations and positively at low. In many species, although the body regions that bear chemoreceptors have only one structural type of sensory hair, reactions to different chemicals vary. The antennae of crayfish, for example, have only one distinguishable type of hair, yet the antennae have distance chemoreceptors functionally resembling those of insects and vertebrates, as well as contact chemoreceptors. This has led some to suggest that there is no differentiation between "taste" and "smell" in these animals, merely differences in thresholds. Nevertheless, the behaviour patterns of crayfish stimulated by different classes of chemicals are different. Receptors in the antennules of a shrimp (*Crangon*) respond electrophysiologically to coumarin (usually considered an odour substance) at concentrations of 0.0001–0.00005 percent, to salt (NaCl) at 1.3–7.2 percent, to acetic acid at 0.01 percent, and to quinine chloride at 0.001–0.0005 percent. The observed differences are sufficient to put coumarin in a separate ("smell" or distance) class from the other (contact or "taste") chemicals, as it is for insects and mammals. Thresholds for the other three substances are on the same order as they are for insects and mammals. Thus, although two structurally different receptors have not been distinguished for crustaceans, these animals still show evidence of two types of chemoreception (distance and contact), as in insects and vertebrates. Perhaps the structural similarity of crustacean antennal hairs masks functional differences in their nerve cells.

Questions of the distinction between "taste" and "smell" in crustacea

Behavioral significance of crustacean chemoreception. Chemically modifiable behaviour patterns are wide-spread among crustaceans and have received considerable study. Feeding responses usually occur in two steps: (1) response to chemicals from food at a distance, mediated through receptors on the antennae, antennules, and sometimes the tips of the legs; and (2) acceptance or rejection upon contact with receptors on the antennae, legs, and mouthparts. Barnacles have receptors that mediate feeding responses when stimulated with glutamic acid, proline, or potassium ions. It is believed that these materials initiate ingestion when they are released from prey that is punctured by spines on the entrapping legs of the barnacles. Electrophysiological studies on specialized appendages (dactyls) of the crab (*Cancer*) show that these respond to a variety of amino acids. Among crabs that feed on fish, the receptors respond to trimethylamine oxide and betaine, both chemicals found in fish flesh.

Parasitic and commensal crustaceans respond to chemicals from their hosts. Receptors on the antennules of commensal shrimps initiate nerve impulses when stimulated with fluid discharges (effluents) from their mollusk or echinoderm hosts. Communication by chemicals within any crustacean species is presumably common in the group but has been little studied. Swimming barnacle larvae aggregate specifically, attracted by a chemical given off by settled (fixed) individuals of the same species. This eventually makes reproduction possible among these fixed animals, since their eggs are fertilized internally. Sperms from one barnacle are transferred by a long penis to a neighbouring individual, this being feasible only because the animals aggregate. Sex pheromones have been reported for certain crabs. When ready to moult to sexual maturity, a female crab (*Portunus*) releases a chemical in her urine that attracts the male. In many species of crabs, the male is attracted from a distance by pheromones but uses his contact chemical sense for final identification of the female before mating.

Reactions to environmental chemicals are almost universal in crustaceans. Intertidal barnacles, like intertidal mollusks, respond when splashed with seawater by opening and becoming active, and they react to fresh water by closing tighter. The receptors that mediate this behaviour

are along the edges of the mantle. Terrestrial isopods (sow bugs) select places that have specific humidities, the preferences varying with species and other environmental conditions. The receptors have been called osmoreceptors (since they conceivably respond to osmotic pressure), but there is no proof that they are distinct from ordinary chemoreceptors.

Insecta. Among the insects, only the blowfly (*Phormia*), the honeybee (*Apis*), and a few species of caterpillars and moths have been given detailed chemoreceptive study. Otherwise studies are scattered, in detail on only one aspect for some species, in others wide-ranging but without detail. Chemoreception in whole orders of insects has been almost entirely neglected; *e.g.*, among Neuroptera (*e.g.*, ant lions), Trichoptera (caddisflies), Odonata (dragonflies), Mecoptera (scorpionflies), and Plecoptera (stoneflies). For *Phormia* and *Apis*, however, investigative evidence rivals that available for man and rat; and understanding of the mechanisms of taste for *Phormia* is better than that for mammals.

Locations and structure of insect chemoreceptors. There is general agreement as to the parts of the insect body that bear chemoreceptors. Distance chemoreceptors are usual on the antennae and on the palpi of the mouthparts. For most insects, the antennae are probably the major locations of these receptors. In the honeybee, each antenna has about 500,000 receptor cells, most of them probably chemoreceptive, the remainder being mechanoreceptive (for tactile stimuli) and thermoreceptive (for temperature). Contact chemoreceptors are on the following structures: external mouthparts, pharyngeal wall (inner mouth), and ovipositor (egg-laying organ) in both chewing and sucking insects; tarsi (feet) and antennae in sucking species. A form of common chemical sense has been reported for insects but has been poorly studied. The receptors seem to be generally distributed over the animal's body, but they are still unidentified.

Chemoreceptors on insect feet

Regions of the insect body known to bear chemoreceptors have many types of so-called hair sensilla, named on the basis of their shape (Figure 9). The following types of sensilla are known from critical behavioral or electrophysiological studies to be chemoreceptive: (1) trichodea (hairs), distance and contact reception; (2) basiconica (pegs), distance and contact; (3) coeloconica (pegs in pits), distance; and (4) placodea (pore-plates), distance.

The following types of structures are suspected of being chemoreceptive: (1) sensilla ampullacea (flasklike pits), distance; (2) sensory patches in the pharynx, contact; and (3) free nerve endings in hairs and integument, common chemical sense.

The shapes of the sensilla are not fully reliable indicators of function. Trichoid sensilla, particularly, are active not only in both distance and contact chemoreception, but also in thermoreception and mechanoreception. Electrophysiological recording of impulses from specific sensilla should help settle the matter. The designations by shape also are not entirely precise, for many types of insect "hairs" are intermediate between typical long thin types and short blunt pegs, and some have extensive modifications of the walls.

In the central cavity of the hair or peg, chemoreceptive sensilla have terminal strands from neuron cell bodies at the base of the sensillum. The nerve cells are usually few in number, and their terminal strands (dendrites) branch variously to lead eventually to micropores (detectable only by electron microscopy) in the walls of the hair or peg. The taste hairs (labellar hairs) on the end of the extensible proboscis of the blowfly (*Phormia*) have been studied most thoroughly. Each of these has three to five neurons that send their dendrites to the micropores, plus a mechanoreceptive neuron with its dendrite attached to the base of the hair. The discovery of these micropores (formerly the exoskeleton of insects was thought to be imperforate) has necessitated considerable reinterpretation of experimental results.

Insect chemoreceptive processes. In the physiology of chemoreception among insects, many types of studies have been made—unfortunately, however, usually scattered among different species. Behavioral studies of feeding responses and other reactions to chemical substances at a distance and in contact, coupled with experimental removal of body parts and similar manipulations, have produced a large published literature. A few insects have been trained to give special reactions to chemical stimuli, the honeybee having been most extensively conditioned chemoreceptively. Some beetles, wasps, ants, flies, and cockroaches have also been studied in this way. Nerve impulses induced by chemical stimulation of the labellar hairs of *Phormia* have been detected electrophysiologically, representing the first time (1955) that a chemoreceptor of any animal was so studied. Since then, electrophysiological studies have been numerous, but mostly with relatively few species of Diptera (true flies) and Lepidoptera (moths and butterflies).

Among selected examples from the history of research on the functions of insect chemoreceptors, studies before 1950 had shown that the principal loci of distance ("olfactory") chemoreceptors are the antennae and that the end organs (terminal structures) are basiconic sensilla and pore-plates. Determinations of response thresholds, differing with the testing conditions, showed that the classes of chemicals to which insects respond at a distance are about the same as those that elicit responses from vertebrates. (The thresholds for series of chemicals are in the same general order for both groups of animals, although absolute values often differ widely.) Some species of insects are found to have distance chemoreceptors on structures other than the antennae, mainly the palpi of the mouthparts. The exact receptors and their properties were little understood in the 1950s.

Since about 1960, electrophysiological studies have yielded major data about the distance chemoreceptors

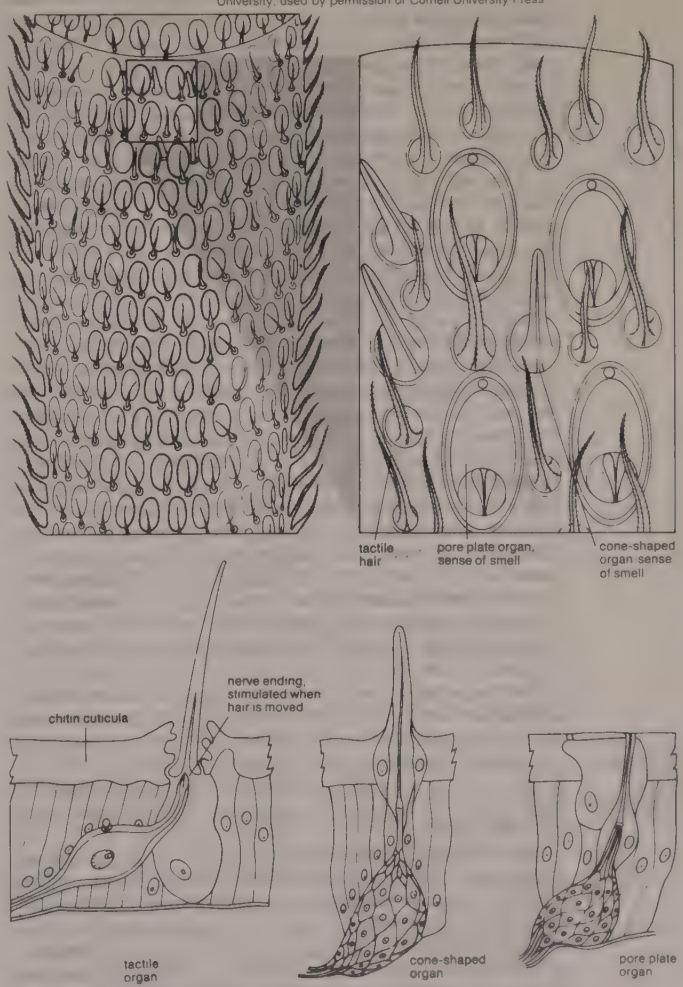


Figure 9: (Top) Segment of worker honeybee antenna; (bottom) cross sections of the antenna's sensory organs

Electrical activity of silkworm antennae

of insects. Nerve impulses are recorded from the antennal nerves to produce so-called electroantennograms. The major species studied are silkworm moths, both the commercial silkworm (*Bombyx mori*) and the giant silkworms (Saturniidae). Males of these species find their prospective mates by means of a special scent given off by the females; receptors on the antennae of males are remarkably sensitive to these special compounds. *Bombyx* males have about 40,000 sex-odour receptor cells on each antenna, with endings in various hairs and pegs. These structures are generally tuned to specific odours and so are called "odour specialists." They can be stimulated with odorant concentrations as low as 100 molecules of the given chemical per cubic centimetre of air. Females of the *Bombyx* species have distance chemoreceptors that are not so tuned; instead, the cells respond to a wide variety of chemicals, being called "odour generalists." The generalist type of receptor cell can respond both by increased neural firing (excitation) or by decreasing firing (inhibition).

Among caterpillars that feed on plants, odours are detected by similar sensilla on the short antennae. These structures are generalists, each responding to a variety of compounds. Their responses, however, differ in a number of ways: (1) latency, the time needed for response after a stimulus is presented; (2) rate of increase in frequency of firing; (3) rate of adaptation, such as loss of responsive capacity as stimulation continues; and (4) alternation of increase and decrease in the frequency of neural firing. Although there are only a few receptors present in the antennae of such a caterpillar, distinctive patterns of these four modes of response to different compounds represent a kind of code that the central nervous system of the animal seems to interpret as, at least, acceptable or unacceptable chemical stimulation.

After many years of behavioral studies on contact chemoreceptive processes among insects, electrophysiological methods have dominated the field since 1955. In many cases these continue to be supplemented by corresponding behavioral observations. The blowfly (*Phormia*) has become the "standard" subject, just as the fruit fly (*Drosophila*) has served in genetics. The labellar hairs of *Phormia* are known to be contact chemoreceptors; when its tip is inserted into a capillary tube containing a sapid solution, the hair responds with electrical changes that may be picked up through the solution. Thus the animal's responses to specific chemical substances can be readily monitored. An extensive mass of data has been gathered with this fruitful system.

Besides having a mechanoreceptive cell at its base, the blowfly's labellar hair has dendrites from four or five sensory cells. Each of these makes electrical responses that distinguish the cell as one of at least four types: (1) salt receptor (or cation receptor), once called L fibre because it produces large spikelike patterns of electrical activity on the recording screen; this cell is stimulated by positively charged ions (cations such as Na^+) and by acids and mediates behavioral rejection in water-satiated flies; (2) anion receptor, stimulated by negatively charged ions, and mediating rejection under all circumstances; (3) water receptor, once called W fibre; this structure fires when stimulated by water and mediates its acceptance by the animal; and (4) sugar receptor, once called S fibre because of its small electrical spike; stimulated by sugars, it mediates their acceptance by the fly.

Thus, rejection or acceptance of sapid solutions largely depends on the blowfly's receptors. A sugar solution causes one set of receptors to fire to bring about extension of the animal's proboscis and to stimulate feeding activity. A solution containing salt or acid stimulates another set of receptors to fire to inhibit extension of the proboscis and of feeding behaviour.

Extension of the proboscis

The stimulating thresholds for a great number of chemicals have been determined with the blowfly, and some general rules have been propounded. The stimulative effectiveness of cations and anions is proportional to the effective intensity of the electrical field generated by the given ion. At least for cations, stimulative effectiveness also seems correlated with the speed at which they move in solution (*i.e.*, their ionic mobilities). The data suggest that the receptor is stimulated by penetration or adsorption of the chemical on the surface; so far neither ionic mobility nor electrical field has been shown to be the only factor that affects thresholds. Rejection of alcohols and of other organic compounds by blowflies seems to be mediated by inhibition of the animal's sugar receptors. Stimulative effectiveness increases with carbon-chain length in a given series of chemicals up to about 11 carbon atoms. The effectiveness seems best correlated with the comparative solubility of the substance in water and oil, suggesting that penetration of the receptor surface is involved in stimulation. The effectiveness of sugars shows no obvious relationship to any of their chemical or physical properties but loosely seems to depend on their nutritional utility to the insect. Lactose, one sugar that is not adequate for nourishing flies, for example, does not stimulate the sugar receptor. Most stimulating are fructose, sucrose, and glucose, in that order; this is the order of their sweetness as tasted by man. In spite of the large amount of data available, however, neither the exact mechanisms of stimulation nor the details of their interrelationships has been worked out for these insects.

Among the so-called pseudotracheae ("false air ducts") on the labellar pads at the tip of the proboscis of these flies are short peglike sensilla (the interpseudotracheal papillae). Studied electrophysiologically, the papillae show evidence of bearing four kinds of receptors: (1) a mechanoreceptor; (2) a sugar receptor; (3) a salt receptor having other sensitivity as well; and (4) one with chemosensory function unknown as yet, although some data suggest that it may respond to amino acids. Specifically, the labellar hairs do not respond to amino acids, yet amino acids are ingested by blowflies.

The electrophysiological activity of taste receptors in *Phormia* has been correlated with the feeding behaviour of the animal. Attraction to foods from a distance is olfactory, mediated by receptors on the fly's antennae and palpi. Extension of the proboscis (at rest it is folded into the head capsule) is brought about by stimulation of sugar receptors, usually in tarsal hairs, sometimes in labellar hairs. Extension can be inhibited by appropriate stimulation of other sensory fibres by salts, acids, or repellent organic compounds. Stimulation of the labellar sugar receptors brings about sucking as long as stimulation of the other fibres is not too intense or provided that inhibition by other organic substances is not too great. Feeding behaviour is maintained and its level of activity is determined by stimulation of labellar and interpseudotracheal sugar receptors. The higher the concentration of sugar in solution, the more avid the fly becomes and the longer it feeds. As feeding proceeds, the sugar receptors adapt to stimulation, finally no longer firing above their resting levels, and feeding ceases. After this, chemoreceptors in the blowfly's foregut take over and shut off feeding behaviour until the meal is moved out. How widely this *Phormia* scheme will be found to operate remains to be seen, but, as studied so far, it seems generally to hold for other insects.

Behavioral significance of insect chemoreception. Studies on feeding behaviour among insects are extensive. Some insects are strictly monophagous (eating only one food); at the other extreme there are highly polyphagous insects (that eat almost any organic matter). Most insects, however, fall between these rare extremes, showing restricted food preferences that depend on the presence of specific marker chemicals (feeding stimulants) in acceptable items of diet.

Insects engage in a tremendous variety of mutual and commensal relationships; to do so they must find symbiotic partners. Many cases of chemical orientation to partners are recorded, usually in connection with the important communication signals of insects. Host finding by insects that parasitize other animals is likewise influenced or determined by chemical signals. Mosquitoes, for instance, find suitable hosts (*e.g.*, human picnickers) by sensing lactic acid, carbon dioxide, and moisture on the victim's skin, as well as by detecting his body heat and movement.

Prey location by mosquitoes

Chemical communication is probably universal among

insects; it is certainly of major importance for the largest and best known groups. The possible practical use by man of sexual communication chemicals (pheromones) produced by insects (or made synthetically) in the control of these animals has led to extensive studies of materials that induce their sexual behaviour.

Social insects (*e.g.*, termites, bees, wasps, and ants) have been known for some time to use chemicals to scent the nest and to recognize individual members of the community. Advances in chemical analysis have facilitated the isolation and identification of many of these compounds. Some of these undoubtedly affect more than the insect's transient behaviour; the so-called queen substance of honeybees (trans-9-oxy-2-decenoic acid), for instance, suppresses development of ovaries in worker bees, often producing (when the swarm is not too large) a community with only one functional female. Similar chemicals are also used for trail marking and as guidance marks to food sources. In ants and stingless bees, deposits of secretions from the mandibular ("jaw") glands (containing such chemicals as geraniol, citral, various terpenes, and methyl ketones) function as guidance spots in the environment to direct fellows to food sources. The most thoroughly studied pheromones of insects are those used for sexual attraction and activation. Specific sexual attractants have been identified in about 250 species of insects. All but about 60 of these are Lepidoptera (moths and butterflies); most of the others are Coleoptera (beetles and weevils). In about 200 of these species, females attract males, and, in about 50 species, males attract females. Generally the attractant substances are what chemists call substituted hydrocarbons, with chain lengths of between eight and 17 carbon atoms in the molecule. It has been theorized that molecules that will allow sufficient structural variety while still being stimulating to insects should have chains of 10 to 17 carbon atoms and molecular weights of 180 to 300. Most of the active substances studied thus far fall within these limits.

Synthetic chemicals that act like the natural pheromones have been prepared for many insect species; these are mainly acetates with chains of 12–16 carbon atoms. Reactions of insects' olfactory receptors to these materials are remarkably specific. In field tests, male moths distinguished the specific chemicals of their own females when these substances were mixed with 26 other pheromones from different species of moths. In the laboratory, where concentrations may be made much higher than in the field, males may confuse some of the compounds, but not under natural conditions. Small differences in molecular structure or configuration can be highly significant. One molecular mirror image (trans isomer) of the Propylure molecule, a substance that attracts pink bollworm males, is active; the other mirror image (the cis isomer) does not attract, yet it masks the trans form when mixed with it.

Remarkably small concentrations of these pheromones can elicit behavioral responses. What was once thought to be the gypsy moth pheromone (isolated in tiny quantities from an extract of hundreds of thousands of female moths) and its synthetic version (Gyplure) have now been found to be inactive in themselves. The active principle seems to be some still unknown impurity present in even more minuscule amounts in the original extracts.

Use of pheromones in insect control

Insect pheromones are thought to be excellent prospects for pest control because of their attractant properties. Unfortunately, most attract males, and even a few fertilized females can maintain a population. At present, the major use of these materials is in population sampling; for instance, male cotton boll weevils (which emit substances called terpenoids) are used in traps to attract females in making a census of their population.

The use of pheromones in insect control is complicated by the finding that high concentrations repel and low concentrations attract. Thus, if high concentrations are used in insect traps to get wide coverage, the animals may be repelled when they get near. Furthermore, a pheromone used in baiting a trap must compete with the attractant from living members of the species. Many pheromones have multiple effects, depending not only upon their concentrations but on environmental factors as well. The so-called Nassonoff gland pheromone of honeybees, for example, consisting mainly of terpenes, serves the insects for attracting workers and queens, for marking food sources, in marking the hive, in scenting prospective hive locations by scouts, and in gathering swarms in flight. Thus, different behavioral reactions to the same pheromone can occur under different circumstances.

As a possible way out of many difficulties, it has been suggested that pheromones could be used to flood given locations with odour. This could fatigue the chemoreceptors of the insects and prevent them from finding mates; their sexual communication channel would be jammed. So far, the few tests of this idea that have been made in the field have not yielded very promising results. Except for short-term, geographically restricted effects, as among insects that live in warehouses where farm products are stored, pheromones for insect control have yet to fulfill earlier optimistic expectations.

Besides responding to food and communication odours, insects are oriented by a variety of other environmental chemical factors. Humidity responses have been extensively studied, but whether the receptors react to water vapour or are hygroreceptors (responding to lack of water) is much debated, with no general agreement. Places for laying eggs are selected by many insects (*e.g.*, mosquitoes and parasitic wasps) by chemical sampling of the prospective sites. Some plant chemicals and a number of synthetic materials repel various insects. There seems to be no generally occurring repellent for all insects, nor has any special relationship between chemical composition and olfactory repellency been discovered.

Protection of man and other mammals from attack by mosquitoes, fleas, ticks (which are arachnids, not insects), and other bloodsucking arthropods has been sought in chemical repellents. Tens of thousands of organic compounds have been tested as insect repellents, mainly for use against mosquitoes. Besides repelling at adequate levels when put on a part of the body that attracts the pests, the compound should not irritate the skin nor be otherwise harmful and should have a reasonable rate of evaporation. In the face of such criteria, few practical repellents have been found. Among those in common use are such substances as dimethyl phthalate, Indalone, Rutgers-612, benzyl benzoate, and Deet; the last is widely used, since it repels many arthropods—mosquitoes, fleas, and ticks. Repellent substances also have been sought among the many warning and alarm chemicals produced by insects, but most of these prove to be irritating to the skin or nose of mammals.

Alarm signals among ants

Alarm pheromones have been studied most intensively in ants, which produce them with special glands to alert their colonies to invaders or to other dangers. The active materials are generally related to hydrocarbons, often ketones; citral and its relatives are important components. Some of these chemicals are also constituents of social and sexual pheromones. Honeybees produce an alarm scent that contains citral and isoamylacetate, among other materials. Formic acid, produced by specialized glands of ants, is found to excite both ants and bees. All of these materials function to alert members of an insect colony when the community is threatened. Other insects (*e.g.*, some beetles) produce strongly repellent chemicals that serve to ward off predators. These chemicals range from apparently harmless but strongly odorous substances to such toxic materials as hydrocyanic acid gas. Among bombardier beetles the ejected spray is even heated by chemical action to about the boiling point of water.

CHEMORECEPTION IN THE VERTEBRATES

Besides the familiar vertebrates (animals with backbones), the phylum Chordata includes some smaller creatures sometimes called protochordates. Little indeed is known about chemoreception in such protochordates (*e.g.*, the lancelets and tunicates) beyond that they seem to show some selection of food and location and that they respond negatively to a variety of foreign chemicals. A group of what are commonly called lower vertebrates is the cyclostomes, such round-mouthed aquatic forms as lampreys and hagfish. Cyclostomes have a well-developed nasal

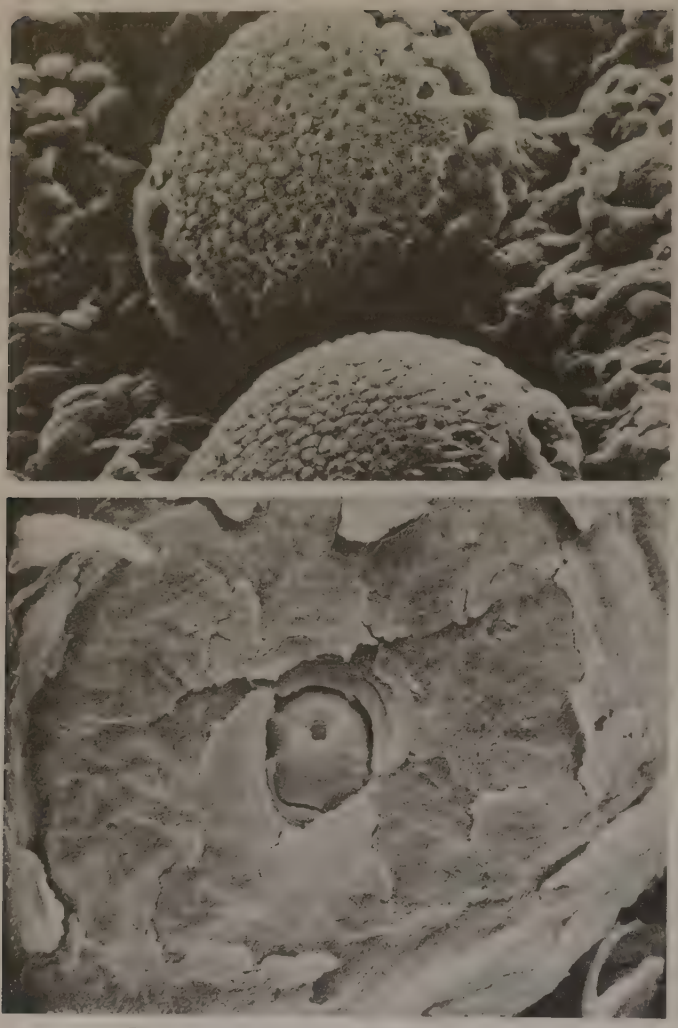


Figure 10: Scanning electron micrographs of (top) two frog fungiform papillae (magnified about 515 ×), and (bottom) taste bud with pore projecting through surface of rat fungiform papilla (magnified about 850 ×).
(Top) P. Graziadei, (bottom) L.M. Beidler, Florida State University

tract, with a single median (central) nostril; they can locate their prey by smell, but otherwise almost nothing is established about their chemical senses. For this reason, the bulk of attention given here to chordate chemoreception will be confined to the five main divisions of vertebrates: fish, amphibians, reptiles, birds, and mammals.

General vertebrate chemoreception. *Gustatory receptors.* The taste buds of vertebrates are secondary sense organs (*i.e.*, sensilla) derived from epithelial cells (Figure 10). Their structure has been well studied by electron microscopy, but in relatively few species (mostly mammals). Each vertebrate taste bud seems to consist of a number of cells of three or four types, but there is some debate as to their exact classification. One widely held view is that the taste bud has four types of cells: so-called supporting cells, sensory cells (the true receptors), basal cells (supplying replacements for old sensory cells), and another type of unknown function. Attempts have been made to designate developmental stages of these types and to view some of them as stages in the development of others, thus giving rise to at least five classes. The sensory cells are continually replaced, each cell having an average life span (at least for rat, mouse, and rabbit) of about 10 days (Figure 11). Each taste bud is innervated by up to 50 nerve fibres entering from below and branching into 200 or more branches to form a basket-like set of dendrites. Presumably chemical stimuli produce electrical changes in the sensory cells of the taste bud, these activating the afferent neurons nearby to generate nerve impulses.

Taste buds of reptiles, birds, and mammals are confined mainly to the upper surface of the tongue, with a few on the pharyngeal walls. In amphibians (*e.g.*, frogs) they are more numerous on the pharyngeal walls and present also on the cheeks and lips. In fish, taste buds are present also on the fins and in some species on the tail. In all cases, vertebrate taste buds are innervated from cranial nerves, mostly the facial and the glossopharyngeal.

Olfactory receptors. Among vertebrates these are the cells of the olfactory epithelium in the nasal cavities. They are primary receptors, true nerve cells the fibres of which form the olfactory nerve leading to the lobe of the brain that mediates the sense of smell. The structure of the cells of this epithelium, as seen with an ordinary (light-wave) microscope, appears remarkably similar for all vertebrates. Electron microscope studies reveal much more structural detail but have not changed the general interpretations. There are three fundamental cell types in the olfactory membrane: receptor cells, supporting cells, and basal cells; in addition, numerous gland cells furnish a mucous covering for the epithelium. Ramifying (branching) among the cells are very delicate terminal fibres of neurons leading to the brain through the trigeminalnerve. These are thought to be receptors of the common chemical sense, responding chiefly to irritants. The olfactory receptor cells have terminal cilia, which are fused into olfactory rods projecting outward.

Olfactory cell types

Man has about 40,000 sensory cells per square millimetre (26,000,000 per square inch) of olfactory epithelium, while the rabbit has about 120,000 per square millimetre, with an estimated total of 100,000,000 such cells. (Fish average between 45,000 and 95,000 per square millimetre, the eel having a total of about 800,000.) A significant discovery made with the electron microscope is that the olfactory sensory cells seem to be synaptically related. Such an arrangement would permit the cells to interact through mutual excitation and inhibition, thus allowing versatility of response at the receptor level itself.

The olfactory epithelium forms at least one wall of the nasal cavity of vertebrates. In fish, the nasal cavities are mostly paired pits or tubes just in front of the eyes, each with two nostrils, one anterior, the other posterior. In terrestrial vertebrates, the paired nasal cavities have external openings, the nostrils (external nares), and paired or unpaired internal openings (internal nares) into the mouth or pharynx. In all cases, water or air is moved through the nasal cavity and over the olfactory epithelium.

Another olfactory receptor of many vertebrates is the so-

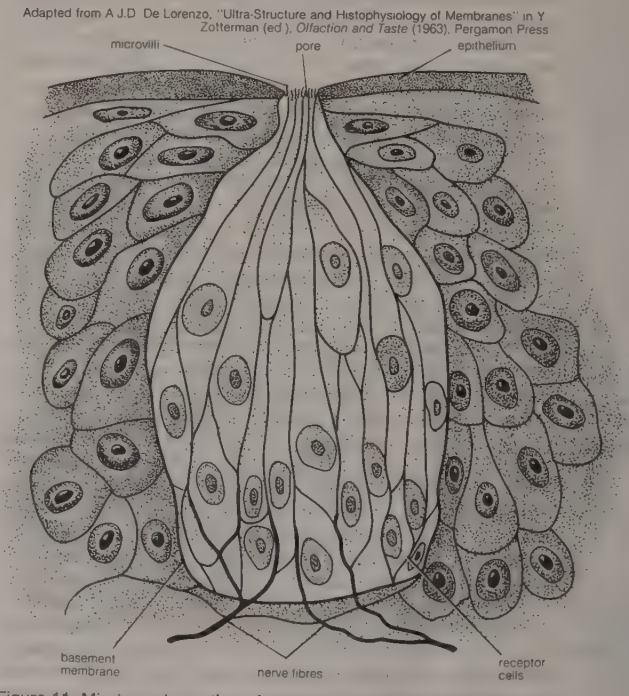


Figure 11: Microscopic section of taste buds of circumvallate papilla.

called Jacobson's organ (vomeronasal organ). This structure is variously developed; absent in fish, birds, and some mammals, it is highly developed in lizards and snakes. Nerve fibres from this organ lead to the accessory olfactory lobe of the brain and so are closely related to the primary olfactory system.

Common chemical receptors. Mucous membranes in vertebrates have receptors that respond to the presence of chemicals rather indiscriminately and, when stimulated, tend to evoke avoidance reactions from the animal. In mammals these common chemical receptors are restricted to the mucous membranes of the nose, mouth, pharynx, eyes, and genital organs. Free nerve endings in the olfactory epithelium of mammals are believed to respond to irritant chemicals.

Free nerve endings

In fish and larval amphibians, free nerve endings all over the animal's body seem to be sensitive to chemicals, their excitation eliciting avoidance reactions. These free nerve endings send their fibres to the central nervous system through spinal nerves. The free nerve endings of the head region enter the brain via the trigeminal nerve. These widely responsive receptors are vitally important in enabling the animals to escape from harmful chemicals in the environment, but relatively few studies have been made on them.

Process of gustation (taste). Among vertebrates other than man, the usual types of behavioral studies (*e.g.*, involving feeding responses) have been made, and training or conditioning procedures also have been used. Gustatory thresholds for detection, acceptance, and rejection have been determined. In more recent years electrophysiological techniques have been most numerous. Human reactions to tasted chemicals can be studied by experiments involving recognition of materials and verbal specification of preference or aversion. Aside from man, the animals most studied are frog, monkey, rabbit, rat, and cat; the investigations have focussed on taste qualities and on the action of sapid substances.

During the 19th century it was widely held that there are four primary taste qualities (salt, sweet, sour, and bitter) and that all other gustatory experiences represent combinations of these. Some investigators have added to these an alkaline and a metallic taste, but others claim that they are not primary qualities. On the assumption that there are four primary taste qualities, chemicals supposedly exemplifying each of the classes (NaCl for salt, sugars for sweet, acids for sour, and alkaloids for bitter) have been applied to the tongues of man and laboratory animals in attempts to find regions of selective sensitivity or (by electrophysiological tests) to locate different types of taste receptors.

Unfortunately taste buds are compound structures, and their neural connections are complex. At any rate, impulses recorded from nerves, or even from single taste buds, fail to give direct evidence about what the individual receptor cells can do. While recordings can be made by inserting fine wires into individual taste buds, the exact cell sampled is not known. It is clear, however, that vertebrate taste receptor cells are not classifiable as sugar, cation, anion, and water receptors as they are among insects. Some vertebrate cells respond to a fairly narrow range of chemicals, but most do not; those cells that respond to salts may also react to acids and sugars, or even water. Certain regions of the tongue tend to be selectively sensitive (*e.g.*, the tip of the human tongue seems highly responsive to sweet chemicals, but not uniquely so). It is no longer expected that, by studying impulses in single gustatory nerves, specific salt, sweet, sour, and bitter receptor cells will be discovered. It seems that patterns of response (rather than specific receptor activation) set up among the sensory cells on the tongue mediate the different taste sensations in man.

As in the case of insects, there is no general agreement on how sapid substances stimulate vertebrate taste receptors. For related series of organic chemicals, stimulative effectiveness is proportional to carbon-chain length up to some maximum and is also related to the comparative solubility of the substance in water and oil. Among inorganic materials, cations generally seem to have stimulative effects that are proportional to their mobilities, but there is great variability in response to the same ions from one vertebrate species to another. Sweet substances are not chemically definable; at least there is no obvious relation of taste with molecular structure. Although many sugars apparently stimulate the same receptors, man and other mammals often can easily distinguish one sugar from the other. Activation or inhibition of receptor cells occurs upon stimulation with different materials. The idea of four primary taste qualities or senses (modalities) has semantic utility, but to date it has not proved useful to investigators as a central dogma in understanding fundamental mechanisms of taste.

Process of olfaction. Studies of smell reception among vertebrates have been similar to those with taste, with electrophysiological methods dominating modern research. The literature on the subject is large, particularly with respect to man.

While attempts have been made to categorize odours in classes that could be considered primary, they have not produced a generally accepted system. The smallest number of primary odour qualities suggested is four, but more than 30 have been offered by some theorists. Attempts to relate odours to chemical structure or to other generalizable physical characteristics of odorous materials have not succeeded. Studies on mechanisms of stimulation of olfactory cells have similarly given rise only to theories, none generally acceptable.

The most active research on human olfaction is concerned with attempts to link odours, such as those of foods or perfumes, with specific chemical structures. Newer analytical techniques, as with insect pheromones, have facilitated the determination of the chemical composition of odorous materials present in the tiny amounts typical of natural products. By these means, extracts from foods can be separated into components with characteristic odours and chemically identified. From the standpoint of olfactory physiology, these studies emphasize the immense capacity of individual olfactory cells to detect a tremendous variety of chemical materials.

From white bread alone, for example, approximately 70 odorants have been identified, including alcohols, organic acids, esters, aldehydes, and ketones. From coffee, 103 separable volatile compounds have been isolated and many chemically identified; it is estimated that at least

From J A Nicol, *The Biology of Marine Animals*, Sir Isaac Pitman and Sons Limited

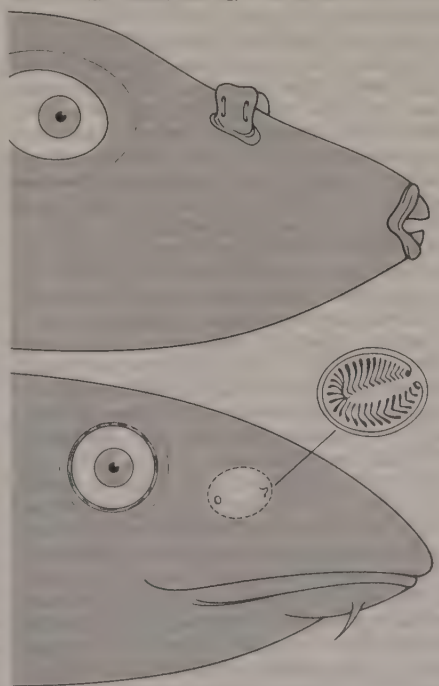


Figure 12: *Nostrils of marine fishes.*
(Top) Puffer (*Tetraodon*) with nostrils on "tentacles."
(Bottom) Cod (*Gadus*) with typical form of nostrils (inset shows detail of folds in nasal cavity).

150 substances contributing to the flavour of coffee will be discovered. Since many of these are present in extremely minute quantities, the capabilities of the human olfactory epithelium, usually regarded as having low sensitivity as compared with that of other mammals, seem remarkable. For substances called mercaptans (*e.g.*, in the skunk odorant), only about 40 receptor cells in the human nose need be stimulated by no more than nine molecules each to give a detectable odour sensation.

Skunk odour

Chemoreception in the main vertebrate divisions. *Fish.* Structure, location, and innervation of fish chemoreceptors are like those of terrestrial animals; thus separation into distance and contact chemoreceptive channels is possible. Taste buds are more widely distributed over the body in fish than in terrestrial vertebrates. In teleosts (*e.g.*, herring, trout, perch) they occur not only in the mouth and pharynx but also on the lips and regions nearby, on whisker-like barbels where present, on fins, and (in some fishes) on the tail. These taste buds are all innervated by branches of the facial nerve. The olfactory epithelium in the fish is in nasal cavities through which water passes; the nasal cavities do not, except in lungfishes, open to the mouth. There is no true Jacobson's organ, although some authors believe that structures near the nostrils may represent a rudiment of this organ (Figure 12).

Feeding behaviour among fish, as with all animals, is determined primarily by the chemical senses, smell being used to find food and taste to determine final palatability. Odours from foods excite movement in hungry fish, but true orientation toward food requires a current to indicate direction.

Social and sexual chemical signals are widespread among fish, though they are probably not as important as visual and possibly acoustic signals. In darkness, or for blindfish in light, species odours are important in schooling. Species differentiation may be excellent; a minnow (*Phoxinus*) can be trained to distinguish 14 different species of fish by their odours, even when the odours are offered in up to 15 different combinations. Mouthbreeders, fishes that hold eggs and young in the mouth, are able to distinguish their own offspring from those of others by odour. Some fish chemically mark their nests with mucus. Redfin shiners, fishes that lay their eggs in the nests of green sunfish, find these nests by the sunfish odour.

Some fish have remarkable powers of olfactory orientation to specific geographic locations. Minnows can distinguish the galaxy of odours of aquatic plants in their home streams and return to them when displaced. The most noteworthy of homing fish are salmon and eels, which return to the fresh water (where they began life) after some years in the sea. Each fish returns to the precise stream in which it was hatched. Many experiments have shown that this is possible only because they sense the odour of the natal stream. Apparently a form of learning called imprinting occurs in these baby fish. The hatchlings learn (or are imprinted) to associate the particular odour of a specific stream with home base. Orientation to the mouths of the streams from the sea requires some other talents as well; but, once the fish enters its home river, it unerringly finds its way to the headwaters where it started life.

Among the earliest reports of animal warning odours was that of the so-called *Schreckstoff* (German for "fright substance") given off by agitated fish. Injured fish produce chemicals that alarm other members of their own species, generally causing them to flee. The material is detectable to fish at extremely low concentrations. Some predators have turned this to their advantage; for example, sharks can detect the odour of an injured fish and swim toward it. In the hope that some chemicals besides those naturally occurring could repel sharks from swimmers, considerable effort has been expended to try to find a suitable shark repellent. So far the results have not been promising, but compounds that are remarkably effective in stimulating other fishes have been found; for example, phenacyl bromide repels teleosts at 0.01 part per million, but unfortunately it does not do this to sharks.

Shark repellents

Amphibians. In spite of the widespread use of frogs in physiology laboratories, understanding of chemoreception in amphibians is extremely meagre, particularly as related to their normal life. The gustatory organs are typical taste buds not only on the tongue and walls of the mouth and pharynx but also variously distributed on the lips. The nasal cavities of urodeles (*e.g.*, salamanders, newts) are relatively simple, but those of anurans (toads and frogs) are complex, with three chambers. The olfactory epithelium is of the usual type; in *Triton*, a salamander that lives both in water and air, cells of the olfactory epithelium have long cilia when the animal is an air breather and short cilia when it is a water breather. The Apoda, wormlike amphibians that are blind, have well-developed nasal cavities and olfactory epithelium. The organ of Jacobson in urodeles is a mere grooved channel in the nasal cavity, but in anurans it forms one chamber of the three nasal cavities.

As usual, feeding is chemically mediated, at least in part, although anurans mostly use their eyes for food capture. There have been few studies on chemicals that determine feeding among amphibians. Chemical communication is probably dominant in salamanders, odours of females attracting males in many species. Females of aquatic salamanders are induced to mate by chemicals produced by males. The chemicals are wafted toward the females by tail-wagging on the part of the males. Frogs and toads seem not to use chemical signals for communication, relying instead on auditory and possibly visual signals. As among fish, salamanders displaced from their home stream are able to find their way back by chemical sensing. Tadpoles, like fish, produce a *Schreckstoff* when injured, its discharge causing other tadpoles to scatter.

Reptiles. Chemoreception among reptiles has been very poorly studied. The major physiological work has been with turtles, mostly with objectives that are totally unrelated to the normal lives of the animals. The taste buds of the turtle are restricted to the tongue and the walls of the pharynx; the olfactory epithelium is in the nasal cavity. All reptiles but turtles have well-developed nasal cavities, crocodilians having exceedingly complex cavities and accessory sinuses. Jacobson's organ reaches its acme of development in lizards and snakes, where it opens into the anterior part of the mouth. Nearby, the lacrimal ducts (tear ducts) open, thus irrigating Jacobson's organ and possibly aiding in its function. This organ is absent in crocodilians and indistinct in turtles. While there has been considerable argument about the function of Jacobson's organ, it is now generally believed that it acts as a second olfactory organ in snakes and lizards at least. The forked tongue of some snakes can be inserted into the openings (inside the mouth) of Jacobson's organ, thus bringing chemical particles picked up on the tongue into contact with the olfactory epithelium. All snakes cannot do this, however, and apparently in some species the materials are dissolved off the tongue in the secretion from the lacrimal glands and thus brought to the organ.

Jacobson's organ in reptiles

Feeding behaviour among reptiles is probably determined to a large extent by chemical stimuli, but there have been few verifying studies. Some snakes find their prey by using the sense of smell, as is shown when newborn young of some snake species attack objects scented with extracts from the skin of species upon which they prey. The receptors involved in this case are in Jacobson's organ. Some predatory snakes cannot trail their prey when this organ is destroyed.

Communication in lizards, turtles, and crocodilians seems to be mostly by visual signals, although some tortoises have glands that secrete chemicals, which they distribute in their territories. Male snakes track females by detecting an odour on their skin; the male will not court his prospective mate if his nostrils are plugged. Rattlesnakes react defensively to the odour of king snakes; conversely, the predatory king snakes track rattlesnakes by their odour. Snakes seem to be more reactive to olfactory stimuli than are other reptiles; nevertheless, reasonable generalizations about the role of chemoreception in the lives of reptiles can only be expected to come from much more study than these animals have had to date.

Birds. General opinion among ornithologists is that birds are predominantly auditory and visual creatures; certainly among birds these senses are usually well developed.

This opinion, however, has led to a possibly unwarranted lack of interest in chemical reception among birds. There is growing evidence that chemoreceptors are well developed in at least some birds. The receptors are well-known: taste buds on the tongue and olfactory epithelium in a rather uncomplicated nasal cavity. The olfactory lobe of the brain in many birds (*e.g.*, kiwis, albatrosses) is large, suggesting a high degree of olfactory sensitivity. Birds have no organ of Jacobson.

Early observations on birds in the field led to the belief that their chemical senses were poorly developed, or even totally absent. Later studies, though few and scattered, suggest otherwise, however. Birds taste water before drinking or bathing, for instance, and their thresholds for rejection are similar to those for mammals. Indeed, birds drink water containing some chemicals at concentrations higher than those they tolerate in bathing water. The bathing thresholds may be well below thresholds for gustatory stimulation of mammals. The large olfactory brain areas of many birds certainly indicate that older ideas about their being olfactorily impoverished need reexamination. The few recent studies that have been made show that birds do, indeed, have an olfactory sense. Quail, for instance, can be trained easily to respond to odours and apparently can mark feeding locations by scent from their bodies, much as mammals do. The chemical senses and the place of chemoreception in the lives of birds—as well as reptiles and amphibians—deserve much more study than they have received up to now.

Mammals. Chemoreceptively, mammals are the best studied of vertebrates by far, and man is probably the most studied of all, although experimental techniques that can be used with blowflies (*Phormia*) and rats are inappropriate for humans. The taste buds of mammals are mostly on the upper surface of the tongue, on so-called vallate, foliate, and fungiform papillae. Some taste buds are also on the palate and in the walls of the pharynx. The olfactory epithelium lies dorsally in the nasal cavity, which in most mammals is extensive and complicated. Bony structures (conchae) subdivide the nasal passages, and sinuses extend into the bones of the mammalian skull. Aquatic mammals alone have relatively small olfactory areas. Jacobson's organ is absent in aquatic mammals, bats, and primates (*e.g.*, monkeys and humans). In almost all other mammals the organ is in the nasal septum (central dividing wall), being small but functional. In a few groups of mammals, Jacobson's organ opens into the mouth cavity through special (nasopalatine) ducts.

Mammalian feeding behaviour is dominated by the chemical senses; indeed, mammals generally are activated and oriented primarily by chemical stimuli. Food finding usually involves olfaction, and food testing involves gustation and olfaction together. Flavours of foods seem to determine acceptance or rejection in all mammals. (Flavour refers to the combined experience of taste, smell, texture, and temperature.)

Flavour testing

Flavour testing of foods for human use is an important factor in the economics of commercial food processing. Therefore an extensive literature exists on the techniques and results of flavour testing and on the production of synthetic flavouring materials. The gustatory organs supply rather restricted information to the brain, but the olfactory receptors supply a vast set of information. As an example of the wide array of volatile chemicals in foods, strawberries contain at least 35 chemical constituents contributing to their odour. These vary from time to time and with conditions in the same berry; for example, crushing converts some materials present in the intact fruit to other substances. The human olfactory organ easily detects these subtle changes, and responses in the brain are thereby affected.

It seems clear that the most important communication signals of mammals are chemical. Social aggregation and territoriality are guided by marking scents secreted by a variety of special glands in different places on the animals' bodies (*e.g.*, on the flanks, back, belly, and near the anus). The secretions are wiped onto objects or sprayed over terrain or are deposited by discharge of urine and feces at particular locations. Almost all mammals chemically mark their nesting or resting areas and quickly detect intruders. Members of a flock or herd (*e.g.*, of sheep) identify one another mainly by scent, apparently producing not only the species scent but also an odour distinctive of that flock or herd alone. Man's use of incense and perfumes in social and religious activities is probably rooted in the basic mammalian pattern of odour sharing within a group.

Similarly, chemical sexual signals are general among mammals. When their nostrils are plugged, male rhesus monkeys and males of some herbivores (*e.g.*, cattle) show no interest in females in heat. Among mice, the odour of strange males (from other communities of mice) interferes with the normal development of fertilized eggs in females; yet, signs of sexual arousal (estrus) can be induced in female mice and other rodents by the odour of a strange male. In probably all terrestrial mammals, arousal of the estral state in females is in response to odours produced by the male genital glands. While fastidious people often may say that sexual odours do not exist for man, the widespread use of perfumes (which supply masked sexual odours) attests to the importance to man of chemical channels in sexual communication.

Orientation to chemical cues is also general among mammals. Many mammals find water or home territories, even when far from them, by the sense of smell. As with fish, it is probable that the total odour complex from soil and plants of a region is detected by mammals.

Mammalian alarm odours

Alarm odours are part of the general communication system of most mammals. Many herbivores have special glands that release odours that alarm the herd when the animals are frightened. Similarly the odour of blood is repellent to many mammals. Many animals (*e.g.*, skunks) have warning odours that repel prowling predators. The tendency of mammals to discharge feces or urine when frightened is also adaptive, for these may act as olfactory repellents to enemies.

Man seems to be an unusual mammal in his limited use of the sense of smell. Other land mammals use olfactory function as their primary sensory basis for interacting with the environment. The sensitivity demonstrated for the human nose with respect to flavour discrimination suggests that even man relies much more than he realizes on the array of olfactory stimuli reaching him from the environment as sources of information. (The human senses of smell and taste are discussed further below; see *Human sensory reception.*)

THEORIES OF CHEMORECEPTOR ACTION

Attempts to create theoretical concepts to explain the actions of chemicals on chemoreceptors have generally been directed toward answering one, or both, of two questions:

1. What characteristics of chemical molecules are critical in producing responses by receptor cells?
2. What molecular characteristics elicit the experiences of particular tastes and smells?

There is still no generally accepted answer for either of these. The theoretical constructs developed have been somewhat different for taste and smell.

Taste (contact chemoreception). Many kinds of actions of sapid substances at gustatory receptor cells have been postulated, and some evidence has supported each. Unfortunately, much evidence militates against each. The most widely accepted possible mechanisms for stimulation of gustatory receptors are the following: (1) chemical reactions at the cell surface; (2) adsorption of molecules on the cell surface; (3) penetration of substances into the cell; (4) enzymatic reactions at the cell surface; and (5) protein bonding in the cell membrane.

Not all of the many adherents of theory (1) select the same type of chemical reaction at the receptor-cell surface. A few physicochemical models of the theory have been proposed, but none fits all the data. The adsorptive theory (2) is probably most widely believed now; while a wide spectrum of data fits this well, not all of the evidence is explained. The penetration theory (3) is supported by correlations between the oil–water solubility and stimulative effectiveness of sapid substances, but the mechanism seems to be too slow and too long lasting. The enzymatic theory (4) can be made to explain almost any data, if one

just imagines the existence of the right enzymes (yet to be discovered). Nevertheless, the temperature independence of taste stimulation militates strongly against it, for enzymatic reactions are strikingly influenced by temperature. The protein-bonding theory (5) is weakened, as is the penetration theory (3), because these processes would be slow to reverse; otherwise good fits with data can be obtained by postulating the existence of appropriate proteins.

Some correlation between human taste responses and chemical composition has been found for sweet, salty, and sour substances, but the results are much less clear-cut for bitter materials. Most substances do not have one of the four simple tastes, and there are other suggested primary taste qualities, the validity of which has not been settled. Almost all investigators who have studied contact chemoreception in detail have come to doubt the validity of any theory of four primary tastes, at least for mammals. More recent electrophysiological data, although gathered mainly by workers who originally adopted the concept of four primary qualities as their guide, do not support the theory. Individual receptors (except labellar hairs of blowflies) are generally not excited by only one of the four presumed primary categories of sapid compounds. The intergrading of tastes for a large series of chemical compounds and the variety of electrical response patterns of receptors obtained in the laboratory suggest more of a continuum of taste-response patterns in a population of receptor cells than the existence of four specialized receptors for primary gustatory qualities.

Smell (distance chemoreception). Theories of olfactory stimulation are even less satisfactory than are those for taste. The events that have been suggested as occurring at the receptor cell to trigger off an olfactory response include the following: (1) chemical reactions at the cell surface; (2) solution of odorant molecules at the surface, thus altering surface tension; (3) radiant energy from an odorant affecting the cell without actual contact of the chemical molecule with the cell; (4) adsorption of the odorant on the cell surface; (5) effect of molecular internal vibrations (molecular resonance) on some aspect of cellular function; (6) enzymatic reactions; (7) penetration of odorant molecules with disruption of receptor-cell membranes; and (8) effect on an olfactory chemical or pigment within the receptor cell, similar to the effect of light on visual pigments such as visual purple (rhodopsin) within the retina of the eye.

Stages in the olfactory response

As would be expected with this array of olfactory theories (and not all proposed ideas are included), there is even less agreement than in the case of taste. The first, fourth, sixth, and seventh of these theories of smell are similar to their counterparts suggested for taste and have the same strong and weak points. The solution theory (2) seems too slow, particularly in accounting for recovery of olfactory sensitivity after adaptation to an odorant has occurred. There is no good positive evidence for olfactory theories based on radiant energy (3) or on olfactory pigments (8). Neither the molecular-resonance theory (5) nor the penetration theory (7) has even majority acceptance right now, although the idea that adsorption (4) is the critical step in stimulation seems to attract adherents.

Attempts to find and name odour primaries have proved more difficult than in the case of postulated taste primaries. A major stumbling block is that none of the theorized primary odour qualities can be related to specific classes of chemical compounds. The postulated primary odours have received such names as: foul, fruity, ethereal, fragrant, resinous, and burnt. The number varies from as few as four primaries to as many as 32, the most usual number being six or eight.

Some current theories relating olfactory experience to chemical or physical characteristics of odorous materials rely upon some postulated selection of primary odours; others do not. Although the first class of theories is based upon attempts to relate specific odours to particular chemicals, no reasonable correspondence between chemical structure and odour has yet been found. Attempts to correlate solubilities or other physicochemical characteristics with odours have been equally unsuccessful. Because many workers believe that the first step in olfactory excitation is adsorption of odorants on the surface of receptor cells, extensive studies have been made on correlations between odour and adsorptive behaviour of chemical compounds at interfaces between water and lipids (*e.g.*, fats or oils). The correlation is surprisingly good in some cases and poor in others. By changing postulated cell-surface characteristics, good correspondence with experimental data can sometimes be obtained, but the theory then potentially seems to fit any data and therefore is suspect.

Two newer, widely discussed theories are based, at least in part, on molecular shape rather than on chemical structure alone. One theory is based on the assumption that odorant molecules puncture the receptor-cell surface, thus releasing ions, and that the ability to puncture the surface depends not only upon the molecule's chemical properties but also on its shape. The olfactory quality experienced is believed to be the result of differential ability of molecules to puncture the receptor cells, determined by the size and shape of the molecules, and by differential rates of healing of the punctures by the cell. Not enough observational data are available on the fundamental events assumed here to make evaluation possible.

Theory of molecular shape

An alternative theory starts with the postulate that there are only seven primary odours, each of which results from the fitting of molecules of seven specific sizes and shapes into special receptor sockets imagined to exist on the cells. Thus molecules of compounds with a similar odour should have similar size and shape, and proponents of this idea believe that this is so. Others, however, find situations that are inexplicable by this "socket" theory. A most critical objection to this theory is that it is impossible to code the tremendous variety of definable smells with a system of only seven units. This has led some investigators to postulate many more than seven primary odours, separable molecular shapes for all of which have yet to be discovered.

Still another theory (5) of odour qualities starts from observations of high correlations between low-frequency molecular vibrations (resonances) and odours. This theory assumes different primary receptor cells, the number still unknown but probably relatively large. The primaries, in this case, are not postulated ahead of time (a priori). Since the theory depends on experimental evidence for its detailed development, only time will tell how or if the correlations will emerge. It is not assumed that the molecular characteristics being measured (*e.g.*, resonances called Raman spectra) are in themselves the stimulative factors; instead it is theorized that they are accompaniments of molecular energy characteristics that are the actual factors in olfactory stimulation. Thus, the unspecified molecular vibrational characteristics are postulated as acting upon energy-transfer mechanisms in the cell membrane or as determining orientation of odorant molecules on the cell surfaces.

None of these theories of smell at present has wide enough acceptance to be said to be the dominant idea. The general attitude is one of wait and see, while proponents of each gather data. Only further research will decide whether any one of these, or none, fits the observed evidence. Theories of gustatory qualities, starting with widely accepted agreement on primary tastes, and those on olfaction, starting without a generally accepted scheme of primary modalities, have now come to about the same conceptual turning point. (H.W.F.)

Photoreception

Photoreception is the activation of a biological process by means of illumination. Most organisms, including man, respond to visible light; some react to wavelengths of light not seen by man; and still others can react to properties of light not detectable by man, such as polarization (vibration of light waves in a definite pattern). This section is concerned with the sensory processes by which animals detect information carried by light (a detailed discussion of the human eye and its function is to be found below; see *Human vision*).

Light energy is necessary for life on Earth. Green plants

require light for photosynthesis, the process by which water and carbon dioxide are transformed into carbohydrates; plants also show adaptive responses (*e.g.*, germination and flowering) to annual changes in daily light periods. Animals depend on plants for food and thus are indirectly dependent upon photosynthesis. In some animals, response to variations in day length is of great importance in the regulation of annual reproductive cycles. (For additional information about the above responses to light, see PHOTOSYNTHESIS and BEHAVIOUR, ANIMAL: *Stereotyped Response* and *Photoperiodism*.)

Light, the name given to the mediator of the sensation of sight in higher animals, including man, and the equivalent of this sensation in lower animals, is the part of the electromagnetic spectrum that is visible to animals; it includes the range of wavelengths from about 300 nanometres (1 nm = 10^{-6} millimetre) in the near ultraviolet to about 700 nanometres in the deep red (300 nanometres is beyond violet and does not evoke sensation in the human eye).

The entire cell of a unicellular animal such as *Amoeba* may be sensitive to light so that the cell moves toward or away from it. Some unicellular animals (*e.g.*, *Euglena*) have developed a light-sensitive receptor, or eyespot—a region with a lower threshold for light stimulation than occurs in the rest of the cell. Some multicellular animals have photoreceptive cells, or eyespots, scattered in various parts of or throughout the body; those in the outer covering of the earthworm (*Lumbricus*) serve in directional orientation, which involves comparison of light intensities at different directions. Most animals have localized photoreceptors of varying complexity—*e.g.*, the ocellus of certain mollusks and arthropods; the compound eyes of arthropods; and the camera eyes of cephalopods and vertebrates.

Evidence indicates that the eyes of certain insects can make use of the information carried by near ultraviolet wavelengths of light as well as that carried by visible wavelengths; both carry information related to the sensation of colour. The eyes of many invertebrates, such as certain arthropods and mollusks, have evolved in such a way that they can detect polarized light; *i.e.*, it evokes a sensation and provides information used for navigation. Visual sensation in higher organisms is primarily a complex response to the intensity and the spatial and temporal distribution of light on the photosensitive retina (the innermost layer of nervous tissue within the eye). Different eyes, different specialized parts of the same eye, and even the same parts of the same eye vary in their responses to illumination. Both the properties of light and those of the eye are thus important determinants of visual sensation. The great differences in the light-analyzing capacities of animals are reflected in the great diversity of gross structural organization involved in photoreception. The fundamental mechanism of photoreception—photochemical activation of a light-receptive pigment and the primary excitation-initiating process—seems to be similar among most animals, however.

This section deals with the optical properties of eyes, including the basic arrangements for image formation and light detection and the morphology of photoreceptors; the photochemistry of light detection; and the physiological functioning of the receptor cells that initiate nervous activity. These initial processes of photoreception provide information for the neural centres of the retina and higher nervous centres. The neural events involving the higher centres lead to visual perception.

THE OPTICAL PROPERTIES OF EYES

The first active step in vision is the absorption of light by a photosensitive substance, a visual pigment. Various devices within the eye assist vision by directing incoming light to this pigment; *i.e.*, they act as light guides by refracting (bending), reflecting (turning back), or guiding light. The arrangement of the optical structures influences the resolving capability and other basic properties of visual sensation.

Camera eyes. *In vertebrates.* The mammalian eye (Figure 13), somewhat like a camera, has a cornea (a transparent, anterior portion) and a lens; it functions as a dioptric system—*i.e.*, a system in which light rays are refracted so as to focus on the retina; the image projected on the retina is inverted. In the retina the light first passes through several layers of nerve cells before impinging on the photoreceptor cells, called rods and cones. The dimensions and refractive powers of all the optical parts of the human eye are known, making it one of the best understood vertebrate eyes (see below *Human vision*).

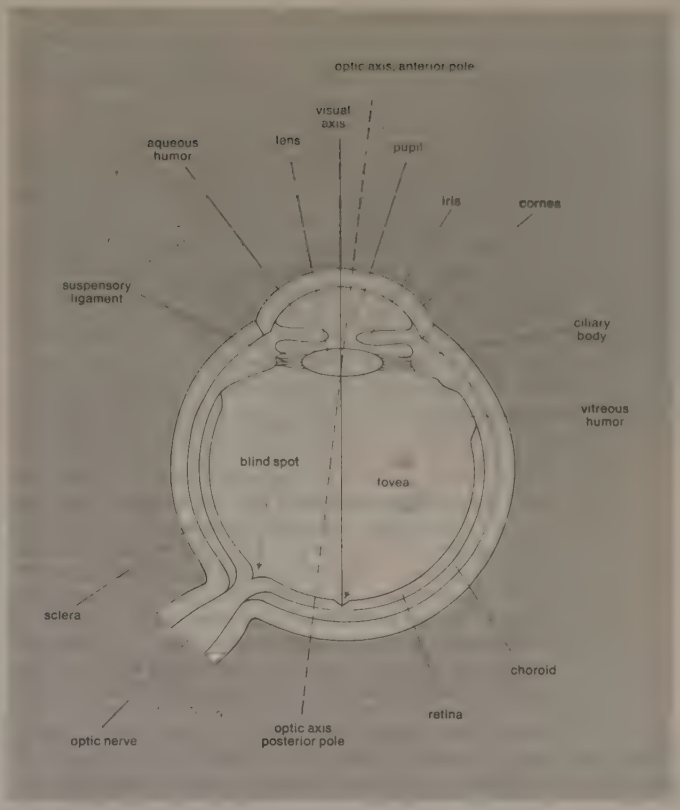


Figure 13: General structure of the mammalian eye.

From P.B. Weisz, *The Science of Zoology*, copyright 1966, used with permission of McGraw-Hill Book Co

The large size of the camera-like vertebrate eye makes it potentially the most efficient of all eyes because it can project a large image on a large surface area containing a high density of receptors. Both vertebrate and invertebrate eyes reflect the influence of the animal's environment. The optical arrangement of the eyes of animals active during the night (*i.e.*, nocturnal) suggests that resolution is sacrificed for light-gathering power (see Figure 14). The opossum lens, for example, is so large that it almost touches the retina—*i.e.*, it has a short focal length, the distance from the centre of the dioptric system to the place at which the image of distant objects is focussed (retina). The short focal length combined with a wide aperture results in a low focal ratio, or f-number

The influence of environment on vertebrate eyes

From G.L. Walls *The Vertebrate Eye*, Cranbrook Institute of Science

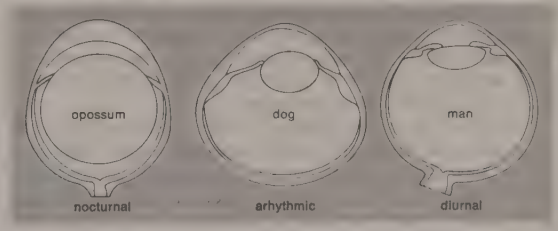


Figure 14: Influence of the environment on the optical arrangements of animal eyes (see text).

$$\left(\frac{\text{focal length}}{\text{aperture diameter}}\right),$$

and ensures high light-gathering ability. In the eyes of animals active during the day (diurnal), the lens is smaller; as a result, the optical centre is closer to the front of the eye.

and its front surface is flatter. Thus, the focal length of the system is longer, the f-number is higher, and the image on the retina is larger and dimmer than in the nocturnal eye. Assuming that the large image can be detected by the photoreceptors, resolution is improved at the expense of the speed of the lens system.

In order to utilize efficiently a large image, the retinas of diurnal animals have localized areas with many photoreceptors; *i.e.*, a higher receptor density. The receptors in this area, called the area centralis, are usually cones, the receptors of daylight and colour vision. These areas for sharp vision, often circular, are seldom located exactly in the optic axis (an imaginary line drawn through the centre of the cornea and the central point of the eye, see Figure 13). The eyes of most birds have two such areas, the centre of each of which is specialized by a thinning of the retina to include only the receptors. This gives rise to a depression called the fovea, also found in teleost fishes, certain reptiles, and man. When the area centralis contains a yellow pigment, it is called the macula lutea. The macula lutea is found in higher primates (simians) and possibly chameleon lizards. This pigment filters out the shorter wavelengths of light and improves the sharpness of the image by reducing the chromatic aberration (variation of the focal length with different wavelengths of light) that would be caused by the inability of the lens to bring the long and short wavelengths to the same focus.

Distant objects are in focus on the retina of the normal human eye. In order for objects closer than about six metres (20 feet) to be in focus, however, an adjustment called accommodation is necessary; otherwise, the image would fall in back of the retina, and the object would appear fuzzy. In mammals, birds, and reptiles other than snakes, the accommodative adjustment consists of sharpening the curvature of the lens so as to shorten its focal length. In snakes, elasmobranchs (*e.g.*, sharks), and amphibians, accommodation is achieved by moving the lens—hence its focal plane—forward. In lampreys and teleost fishes the eye is adjusted for near objects, and accommodation for distant vision is carried out by a backward movement of the lens. Some species have evolved adaptations that make accommodation unnecessary. The retina of the fruit bat (*Pteropus medius*) is in folds, ensuring that some part of it will intercept an image at any location. The ray *Raja batis* and the horse have ramp retinas, in which a continuous and gradual change occurs in the distance between the lens and retina in certain parts of the eye. Specific areas in these animals' eyes are presumably used to view objects at varying distances much as the human eye directs the image for detailed vision onto the fovea.

Accommodation in the eyes of vertebrates

Because vertebrate species are adapted to almost every aquatic and terrestrial environment, they have evolved equally diverse eyes. In air, for instance, the front surface of the cornea can function effectively for image formation; in underwater eyes, however, the refractive index (the ratio of the speed of light in air to that in a given medium) of water and the cornea are almost identical, and the corneal front surface does not refract light. In these eyes the lens does much of the image formation.

Vertebrates have two types of photosensitive cells, rod cells and cone cells. The rod cells, which are long and fat, contain large amounts of visual pigment; they are the photosensitive cells for vision under conditions of dim illumination (scotopic vision). The cone cells, which are relatively small, mediate daylight vision (photopic vision) and colour sensation in many animals. The photosensitive photoreceptor outer segments of rods and cones are stacks of disks, or lamellae, with the planes of the disks at right angles to the long axis of the rod and cone cells. The retinas of animals active both day and night, as are those of humans, contain both rods (for night vision) and cones. In parts of the human retina the rods and cones are intermingled; elements of the nervous system provide the switching mechanism that permits adjustment for light conditions. The specialized fovea contains only cone cells; in the fovea the switching function is accomplished by eye muscles that change the direction of the field of vision in order to bring the image to the fovea.

The amount of light reaching the photoreceptor cells is controlled to some extent by the pupil, the opening of the eye through which light passes. The iris, the coloured portion of the eye surrounding the pupil, constitutes a diaphragm. Its muscles cause the pupil to change in diameter, decreasing the size of the pupil when light enters and increasing it when little or no light enters. The area of the pupil increases about 15 times in going from one millimetre to four millimetres (0.04–0.16 inch) in diameter, a relatively small increase in comparison with the range of light intensities under which the eye effectively operates. Since the amount of light entering the eye is proportional to the size of the pupil, it can be seen that changes in pupil size modify the amount of light only over a small range. Changes in pupil size are important in the human eye because they allow the lens to be used most effectively for visual acuity. When the whole lens is used, as in dim light when the pupil is large, the image formed by the lens is rather poor, chiefly because of chromatic aberration. The neural image on the retina is already poor, however, because the responses of thousands of rods must be pooled to obtain maximum sensitivity. Use of the whole lens is beneficial because it adds further light without reducing the image quality. When illumination is bright, the pupil is small, and only the aberration-free central part of the lens is used. This high-quality image is used effectively by the cone receptors of the fovea. There, no pooling of receptor responses occurs.

Pupils that form a circle when closed cannot greatly change in area; however, a pupil that forms a slit when closed can close almost completely. When nocturnally active animals find themselves in bright sunlight, they need additional protection for their sensitive rod-containing retinas; such animals have evolved pupils that close to form a slit. Many nocturnal vertebrates also show eyeshine (*e.g.*, the glow of a cat's eyes reflecting light at night). Eyeshine, which is caused by a mirror-like reflection from either the retina or choroid (a layer of blood vessels and connective tissue), enhances the sensitivity of the eye. The reflection of the light outward means that it passes the receptors a second time, giving them a chance to absorb light that was not absorbed during the inward passage through the receptors. Some animals thus have smaller rod receptors than they would otherwise need.

Adaptations of the eyes of nocturnally active animals

In cephalopods. The eyes of the invertebrate cephalopods—octopus, squid, and cuttlefish—are usually cited as examples of convergent evolution because they have independently evolved large camera-like eyes similar to those of vertebrates. The cephalopod eye lies within a cartilaginous cup. It consists of a cornea, lens, iris, and retina with the same basic relations to one another as are found in vertebrate eyes. Iris muscles can enlarge and narrow the pupil. Many details, of course, are different; for example, although the maximum density of photoreceptors in cephalopods is high—about 50,000 per square millimetre (32,000,000 per square inch) in *Loligo* and 100,000 per square millimetre in *Sepia*—the retinal structure otherwise bears little resemblance to that of the vertebrate. The photosensitive cells are of two types and are organized to detect polarized light (see below *Morphological features*). In addition, unlike vertebrate receptors, those of the cephalopod are the first element of the retina to be illuminated (rather than the last, as in vertebrates), and the optic ganglion (a mass of nerve tissue) is separate from the retina (rather than an intimate component, as in vertebrates). Both vertebrate and cephalopod retinas show pigment migration and movement of the receptors in response to adaptation for conditions of light and dark.

The cephalopod cornea does not have any focussing function. Image formation is accomplished entirely by the lens, which is forced forward for viewing nearby objects. The pupil is round in deep-sea cephalopods, such as *Loligo*, and slit-shaped in shallow-water dwellers, such as the octopus. Cephalopod photoreceptors are very long, an adaptation for nocturnal or deep-sea and low-light level conditions. The length allows for the presence of more visual pigment and hence greater absorption of light by each receptor cell.

Eyespots. Eyespots, the most primitive eyes, are found in the protozoan flagellates (unicellular animals with a

flagellum, or whiplike structure, used for locomotion), flatworms (Platyhelminthes), and segmented worms (Annelida). An eyespot may be a specialized part of a cell as in protozoans, a single photoreceptor cell, or a small cluster of receptors with few or no accessory optical and neural structures. The entire epithelium (skin) of the annelid earthworm *Lumbricus* contains isolated light-sensitive cells that are eyespots. These photoreceptor cells are rather spherical in shape. A rhabdomere—a structure containing photosensitive pigment—lines a vacuole, or internal cavity, of the photoreceptor cell. The function of the vacuole may be to gather light.

Eyespots in planarians and amphioxus

The flatworm *Planaria* has a more highly developed eyespot. A number of photoreceptor cells are clustered under the epidermis. All of the rhabdomeres, which occur together within a cup-shaped collection of pigment cells, are located on slender filaments some distance from a cell that also gives rise to a nerve fibre.

A third type of eyespot is found in the nerve cord of the cephalochordate amphioxus. Each of a small cluster of photoreceptors, the Hesse cells, has a rhabdomere along the edge that faces a pigmented cell.

Ocelli. The ocellus, which is recognized as a true eye, is similar to a camera in that it usually projects an inverted image onto a light-sensitive layer. The ocellus is distinguished from the compound eye, which has many lenses, and from the more highly developed camera-like eye of the mollusks and vertebrates.

In mollusks. In a simple ocellus, that of *Nautilus,* the photoreceptor cells are bipolar—*i.e.,* the rhabdomere is at one end, the nerve fibre at the other—and arranged in a cup-shaped sheet. There is no lens or cornea, only a pinhole opening.

The more complex ocellus of the slug *Agriolimax reticulatus* is located at the tip of the tentacle (see Figure 15); there is a cornea under the epithelium, a vitreous body

From P.F. Newell and G.E. Newell, *Invertebrate Receptors* (1968), The Zoological Society of London, No. 23

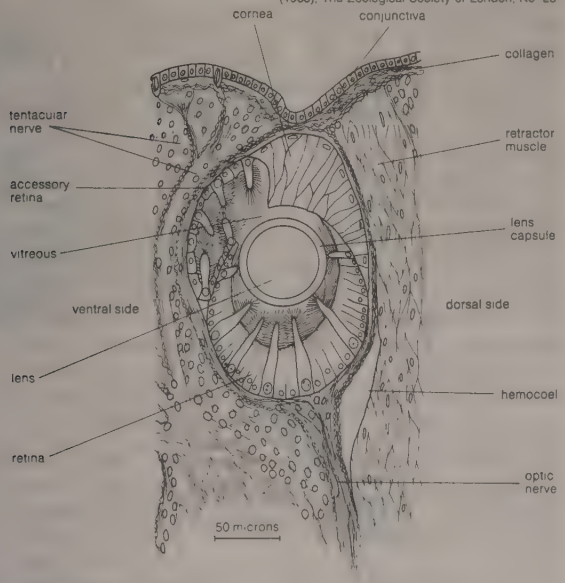


Figure 15: Longitudinal section of the eye of the slug *Agriolimax reticulatus.*

(a mass of clear jellylike material), and a lens, as well as a main retina and an accessory retina. The accessory retina is believed to function as an infrared receptor. As the tentacle is withdrawn, the accessory retina is rotated so that it is exposed to incoming radiation. The few photoreceptor cells are surrounded by pigment-containing cells. The ovoid eye is about 0.18 millimetre (0.007 inch) in its longest diameter. Distant objects appear in focus in the photoreceptor cells of *Agriolimax,* and the shape of the eye changes as the state of retraction of the tentacle varies. The change in shape may provide a mechanism for accommodation, although accommodation may not be particularly useful, because indications are that this ocellus does not clearly distinguish form.

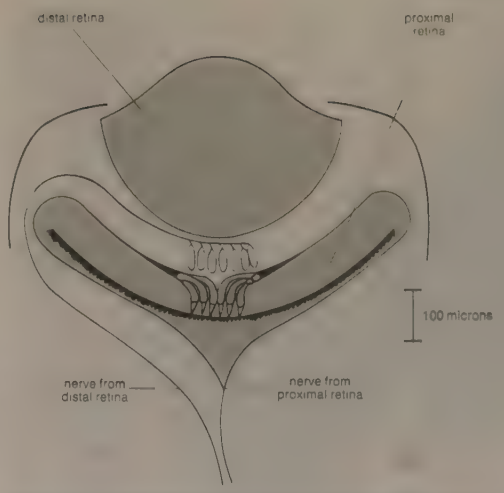


Figure 16: The central region of the eye of the scallop *Pecten.*
From M.F. Land, *Symp. Zool. Soc. London No. 23* (1968), The Zoological Society of London

The molluscan ocelli described above resemble a miniature simple camera, in which an inverted image is projected onto a photosensitive retina. The optics of other ocelli, however, are more sophisticated; for example, the scallop *Pecten* (Figure 16) and related genera have about 100 eyes located along the fringe of the mantle, the lining of the inner surface of the shell. The eyes are one millimetre (0.04 inch) in diameter and have a double retina; one is called a distal retina, the other a proximal retina. The nerve fibres of the distal (*i.e.,* farthest from the body axis) retina respond only to decreases in light intensity; the nerve fibres of the proximal (*i.e.,* closest to the body axis) retina, on the other hand, respond only to increases in light intensity. Because the photoreceptors in the proximal retina are inverted, light passes through all parts of them before reaching the light-sensitive pigment, and the photoreceptors just touch a reflecting structure, the tapetum lucidum. The scallop tapetum contains about 35 layers of thin crystals; the thickness of and the degree of separation between crystals are precise. The crystals and the intervening spaces function as an interference filter. The small reflection from each crystal interface adds to give a large net reflection from the surface of the tapetum. The fact that the photoreceptors just touch the tapetum ensures that they are illuminated on the second pass (after tapetal reflection) by the same light that passed through them the first time. The tapetum in the scallop eye acts as a concave (*i.e.,* depressed toward the centre) mirror, projecting light through the proximal retina and focussing an image of a distant object on the photoreceptors of the distal retina. The combination of tapetum lucidum and inverted photoreceptors in the proximal retina may thus enhance sensitivity without sacrificing resolution. The structural details of the tapetum lucidum have been described not only for the eye of the scallop *Pecten* but also for several arthropod and vertebrate eyes.

The tapetum lucidum

In arthropods. Among arthropods, ocelli are the main organs of sight in arachnids such as spiders and in insect larvae that undergo complete metamorphosis (*i.e.,* a radical physical change during development). Insects that undergo incomplete metamorphosis have three ocelli arranged in a triangle on the dorsal, or top, part of the head; these are subsidiary, however, to the main organs of sight, the compound eyes.

Spiders have two kinds of ocelli, the principal, or anteromedian, eyes and the lateral eyes. The principal eyes of jumping spiders are used when stalking prey and in courtship; the photoreceptors are arranged in four layers in the optic axis (see Figure 17). The two deepest layers cover the entire retina; the two most superficial layers are confined to the retina's central region. The rhabdomeres of the three deepest layers are rod shaped. Those of the most superficial layer are ovoid and oriented in the direction of propagation of light through the eye. Because the receptors are layered, the images of objects at differ-

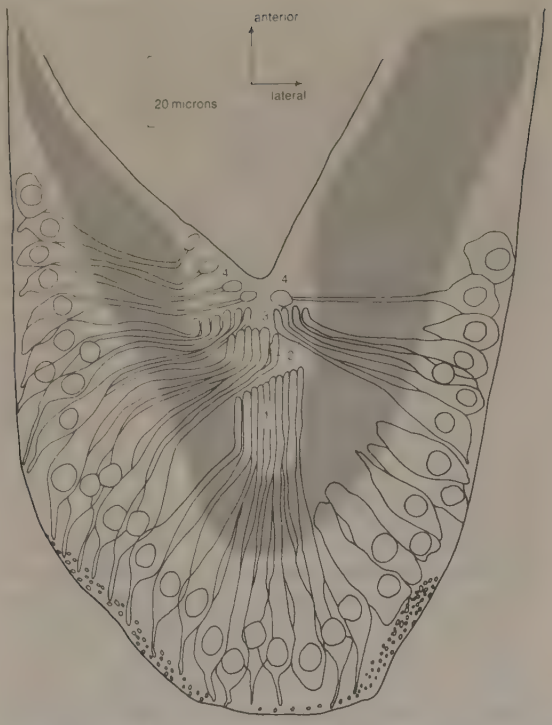


Figure 17: Frontal (horizontal) section close to the centre of the retina of the spider *Metaphiddipus aeneolus*, with the layers of receptor endings numbered 1 to 4. In layers 1, 2, and 3 the presumed receptive part of each receptor is the straight terminal portion; in layer 4 it is the terminal ovoid swelling.
From M F. Land, *Journal of Experimental Biology*, vol. 51 (1969)

ent distances from the eye appear in different layers; it is possible that this type of retinal system is used for depth perception. It appears more likely, however, that the arrangement is used for colour vision. Images of distant red objects appear in the first retinal layer, blue-green images appear in the second layer, and ultraviolet images in the third layer, because the primitive lens system shows considerable chromatic aberration.

No image for visible light appears in the most superficial ovoid rhabdomeres of the principal eyes of spiders, and it has been postulated that they function to detect polarized light. The principal eyes also have a set of six muscles that produce eye movements when an image falls on the retina. The eye follows the image, which remains centrally fixed.

The lateral ocellus of the caterpillar, which has a group of five or six ocelli on each side of the head, resembles one ommatidium, or unit, of a compound eye. The optical apparatus includes a lens, or cornea; a crystalline cone, which is the principal focussing device; and two layers of photoreceptors (distal and proximal) in the optic axis, each of which is radially arranged like sections of an orange. The rhabdomeres of both cell types are in close apposition over much of their length. Objects at different distances project images at different vertical locations in the rhabdom—*i.e.*, a group of rhabdomeres in physical contact with one another. The proximal and distal rhabdoms form an optically continuous structure, a light guide. The optical information in this structure depends on the light accepted at the distal part of the rhabdom, rather than on the locations in the rhabdom at which images are projected, because the rhabdoms actually touch one another and effectively form one optical structure (see *Optical properties of photoreceptors* below).

Spatial resolution in ocelli and camera eyes

No optical system can form a point image of a distant luminous point, such as a star; because of diffraction (a modification in which a redistribution of energy occurs), the image is instead a small spot, the Airy disk, with a bright centre enclosed in concentric, alternately bright and dim rings. The radius of the first dark ring indicates the size of the diffraction pattern. The smaller the aperture and the longer the wavelength, the bigger the Airy disk and the poorer the resolution.

The degree to which images can be detected is determined both by the size of the Airy disk and the size, separation, and density of packing of photoreceptors. Compared with the eyes of most invertebrates, a relatively large number of photoreceptors occurs in the principal eye of the jumping spider. This suggests that the image is good, that its quality is primarily determined by the size of the lens, and that the mosaic of receptor units makes maximal use of the image quality. By comparison, the theoretical resolution of the human eye is 10 times better than that of the spider. This difference results solely from the difference in the size of the eye. In both the human eye and that of the spider, the focal length is adjusted to transmit the image onto a retina containing about the same density of photoreceptors. Actual resolution, however, is complex and depends on other factors, such as the nature of the target, properties of the illumination, the form of measurement, and neural processing in the eye and brain.

In the lateral ocellus of insect larvae, the size and density of the receptors are probably the limiting factors in resolution. In addition to poor spatial resolution, the light-gathering power of the lateral ocellus is also inferior to that of the spider eye.

Compound eyes. Aggregations of ocelli are found in several lower animals—in polychaete worms, for example. True compound eyes are found only in the arthropods, however. The compound eye of insects is composed of hexagonal or rectangular-shaped, closely packed optical units called ommatidia (small eyes); each ommatidium is virtually a single eye. In different species the size, number, and structure of ommatidia vary. An ommatidium (Figure 18) is composed of a corneal lens, or facet, which consists

From *Journal of Cellular and Comparative Physiology* (1965)

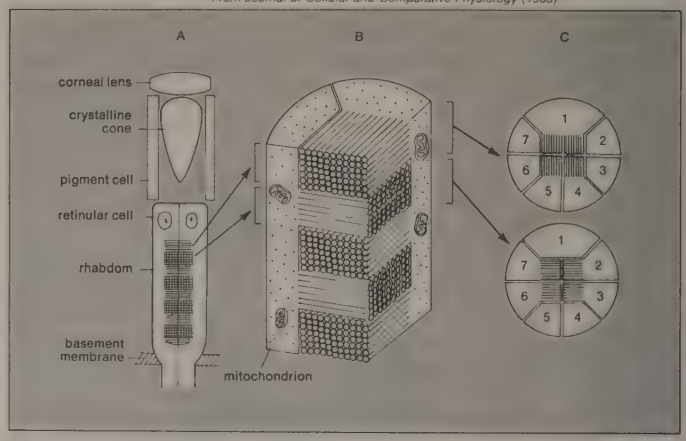


Figure 18: *An ommatidium from the compound eye of the crayfish Procambarus clarkii.*
(A) Longitudinal section through the optic axis with the cone stalk between the crystalline cone and the retinula omitted. (B) Part of the rhabdom. The direction of the closely parallel microvilli in one layer is perpendicular to that in the next. (C) Cross sections through the retinula at the levels of two neighbouring layers, rhabdomeres from retinular cells 1, 4, and 5 constituting one layer (upper figure); 2, 3, 6, and 7 constituting the next layer (lower figure); 1, 4, and 5 the third; and so on alternately.

Components of the ommatidium

of a modified extension of the cuticle (the hard outer covering of arthropods) on the surface of the eye; four cells called Semper's cells or cone cells, which form the crystalline cone; and a sensory region called the retinula (small retina). In primitive insects (*e.g.*, the springtail *Lepisma*), in which the transparent cone cells are not specialized, the ommatidia are called acone ommatidia. In the more common eucone ommatidium, which occurs in moths and butterflies, the cone cells have a more complicated structure and contain granules of glycogen, or animal starch; because the granules are packed at various distances from each other, the refractive index varies in different positions in the cell. In certain beetles (*e.g.*, *Lampyris*) and in the horseshoe crab *Limulus*, the crystalline cone is an extension of the cornea.

The sensory part of the ommatidium, the retinula, consists of several radially arranged cells (retinular cells); each

has a photoreceptor component, or rhabdomere (see Figure 18). The rhabdomeres of neighbouring retinular cells may be either in contact (forming a rhabdom) or completely separate. The optical isolation of each ommatidium is enhanced by its being surrounded by light-screening, pigment-containing cells. During adaptation to light and dark conditions, migration of pigment in cells around the crystalline cone, the corneal process, the proximal part of the ommatidium, and within the retinular cells has been reported to occur in most types of compound eyes.

The apposition eye. In the compound eyes of diurnal arthropods, each ommatidium is separated from its neighbours by pigmented, or iris, cells under all conditions of illumination. As a result, the rhabdom of each ommatidium receives light only through its own corneal lens; light from the lenses of other ommatidia is blocked by the pigment. This is the basic structure of the apposition eye (Figure 19). There are, however, variations in structure.

From G A Mazohkin-Porshnyakov, *Insect Vision* (copyright 1969), Plenum Press

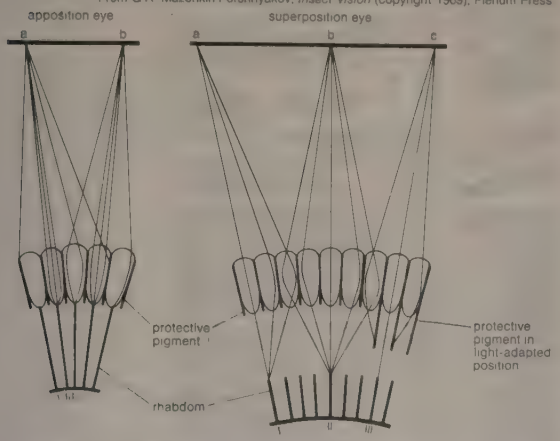


Figure 19: Image formation in apposition and superposition eyes: I, II, III represent different ommatidia; a, b, c represent point objects (see text).

In the honeybee *Apis mellifera,* the cornea is layered, each layer having a different refractive index. It has been shown that the corneal lens projects a small inverted image of a distant object in the crystalline cone some distance short of the rhabdom of an ommatidium. If the image were focussed exactly on the rhabdom, it might effectively transmit image-location information by means of differing responses of the individual retinular cells within an ommatidium. Because the image is focussed in front of the rhabdom, however, individual retinular cells within an ommatidium do not respond to image location. The defocussing also has the effect of broadening the field of view of the ommatidium. As the image moves, it can still be guided into the rhabdom by the crystalline cone, which is considerably wider than the rhabdom.

Specialization of the butterfly eye

It is not possible to generalize from the known optics of the compound eye of the bee to other apposition eyes as, for example, those of the butterfly, which superficially resemble the bee's in having a cylindrical rhabdom. Certain features of the butterfly eye are known, however (see Figure 20). Under the rhabdom of each ommatidium, the tracheole, or respiratory tubule, which supplies air to the ommatidium, has become specialized to form a tapetal filter comparable with that of the *Pecten* eye; it reflects highly saturated colours, usually blue colours in the dorsal ommatidia and more reddish colours in the ventral part of the eye. Each ommatidium thus has its own mirror. (The swallowtails, *Papilio,* are the only butterflies lacking this adaptation.) The light in the rhabdom of an ommatidium is tinted by the coloured light reflected from the mirror, and can be seen only with the aid of an ophthalmoscope—an instrument for viewing the interior of the eye. This filter system probably aids vision by enhancing contrast between objects of certain colours.

Butterflies with a tapetal filter, as well as nocturnal species (with a tapetum of different structure), have a corneal surface structure that acts as an anti-reflector. It consists of minute conical corneal "nipples" that provide a gradual transition between the refractive index of air and that of the cornea, effectively eliminating front-surface reflection for many wavelengths of light. One of the functions of this coating may be to minimize reflection of images from the tapetum lucidum back into the eye.

The rhabdomeres within an ommatidium in the compound eyes of flies (Diptera) have a microvillar structure (*i.e.,* one involving minute hairlike projections; see below *Morphological features*) and are completely separate from one another, in contrast to the apposition eyes above, in which the rhabdomeres are in contact. Each ommatidium contains eight retinular cells and eight rhabdomeres.

From any relatively distant point in space, a small cluster of ommatidia in a fly's eye appears darker than other areas because these ommatidia are the ones that are best aligned to absorb light coming from the direction in which the ommatidia are pointed. This cluster, called the pseudopupil, is often found in insect eyes; it moves, appearing to follow an observer viewing different parts of the compound eye. In the compound eye of the butterfly it is the pseudopupil that lights up with colours when observed and viewed from the same direction as the illumination with the ophthalmoscope. The pseudopupil of flies usually consists of seven ommatidia. In this apposition eye a distant point forms an image on the distal end of the rhabdomere, not within the crystalline cone as in the apposition eye of the bee. From any distant point in space only one rhabdomere in each ommatidium of the pseudopupil is illuminated by a point source; *i.e.,* one rhabdomere in each ommatidium of the pseudopupil is directed toward the same point.

Spatial resolution in the apposition eye

Experimental results suggest that the resolution of the compound eye (*i.e.,* the capability of the eye to distinguish between two separate but adjacent objects) is determined by the divergence angle between the ommatidia. If a compound eye views a pattern of alternating black and white bars in which the angle formed by one bar is the same as that between the ommatidia, the image of one black bar falls on one ommatidium, and the image of the flanking white bars falls on its nearest neighbours. This defines the limit of resolution; if the image of more than one bar falls in an ommatidium, the bars are not resolved.

The image is out of focus at the rhabdom in the bee eye. The angle between fly rhabdomeres is the same as the angle between ommatidia. The fused rhabdom of the bee and the separate rhabdomeres of the fly probably convey

From N Yagi and N Koyama, *The Compound Eye of Lepidoptera Approach from Organic Evolution* (1963) Maruzen Co., Ltd

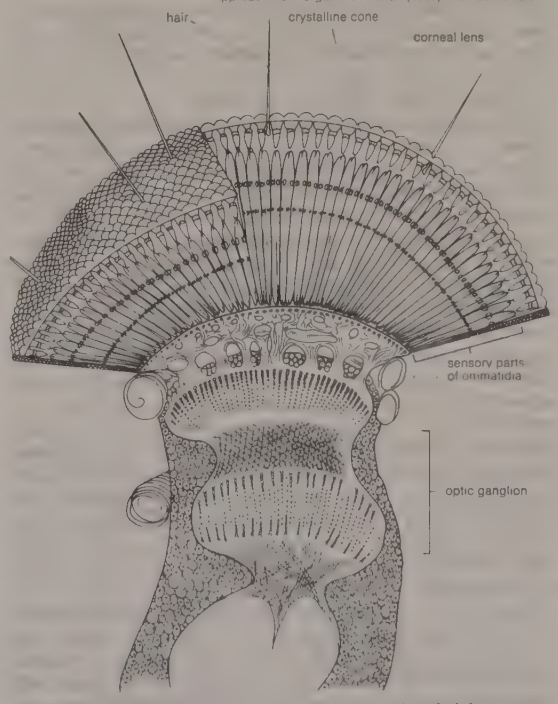


Figure 20: Section through the compound eye and optic lobe of a butterfly.

polarization and colour information rather than spatial information.

The array of corneal lenses on the surface of the eye can be considered as sources of light that interfere deep within the eye to give high-order diffraction images. The structure of the eye suggests that this effect is not detectable by the photoreceptors. Substantial shielding pigment between the ommatidia reduces the intensity of such patterns, which thus should be relatively ineffective in exciting the rhabdom.

The superposition eye. In the compound eyes of nocturnal arthropods, the rhabdoms are deep within the eye, far from the cornea and crystalline cone. In 1891 Sigmund Exner, an Austrian physiologist, reported experiments that showed how these eyes function. He demonstrated that there is pronounced pigment migration within the iris cells. In eyes adapted to darkness, the pigment of these cells migrates upward and into spaces between the crystalline cones of neighbouring ommatidia. In the light-adapted eye, on the other hand, the pigment migrates into the region between the cones and the rhabdom, in effect isolating each ommatidium by surrounding it with a light-absorbing tubular pigmented structure.

The pigment migration has the effect of changing the sensitivity of the eye. When a light-adapted eye is placed in the dark, there is an initial increase in sensitivity of about 10 times with no accompanying pigment movement. In the following 25 minutes in the dark, the sensitivity increases about 100 times, as the pigment gradually withdraws from between the cones.

Exner found that the corneal process of the firefly *Lampyris*, a structure that resembles the crystalline cone of butterflies and other nocturnal insects, has a higher refractive index at the centre than near the periphery and effectively bends incoming light. When the eye is dark-adapted, light from many facets converges in the rhabdoms of many ommatidia forming a so-called superposition image (Figure 20). This eye differs from the apposition eye in that light from many facets is involved in forming an image in the rhabdom of an ommatidium; in the apposition eye, on the other hand, light from its own corneal lens reaches the rhabdom within a particular ommatidium. The mosaic image of the superposition eye, although less sharply defined than that of the apposition eye, is brighter and thus a valuable adaptation for nocturnal insects. In the light-adapted condition, inward migration of pigment in the iris cells effectively prevents the spread of light to adjacent ommatidia, and the superposition eye acts somewhat like the apposition eye; *i.e.*, only the light from one facet reaches the rhabdom within an ommatidium.

Important differences between apposition and superposition eyes

THE PROPERTIES OF PHOTORECEPTORS

The photoreceptor cell absorbs light energy and transforms it into a nervous response. The actual photoreceptor component, or organelle, of the photoreceptor cell contains a coloured substance (visual pigment) that absorbs light and initiates the chain of chemical reactions leading to nervous excitation. The vertebrate photoreceptor cells, the rods and cones, are so called because of the shape of their photoreceptors, which are found in the outer segments of the cells. Invertebrate photoreceptor cells, the photoreceptors of which are the rhabdomeres, show greater diversity of structure than do those of vertebrates.

The function of photoreceptors is dependent on the visual pigment they contain; the identification of photoreceptors ultimately depends, therefore, on correlating the light-absorbing properties of the pigments within them and the physiological response of the receptor cells. Such evidence exists for a number of invertebrates and vertebrates; the identity of some vertebrate and many invertebrate photoreceptors, however, has been inferred from their location, colour, and structure after comparison with known photoreceptors.

When viewed with the light microscope, the outer segments of vertebrate receptors and the rhabdomeres of invertebrates can be recognized by their shape, location, and high refractive index. In the electron microscope, all photoreceptors are seen to consist of a dense collection of membranes. The outer segments of vertebrate photoreceptors are lamellate (layered). Invertebrate rhabdomeres usually consist of microvilli (minute projections of the receptor cell membrane) and resemble a tiny honeycomb in three dimensions. In certain invertebrates, however, the rhabdomeres are lamellate, resembling those of vertebrates. Some features of photoreceptors from various animal groups are discussed below.

Morphological features. *Invertebrate photoreceptors.* The photoreceptor of the jellyfish *Polyorchis penicillatus* (Figure 21) is a modified cilium, or hairlike structure, as are the photoreceptors of many higher animals. Cilia are

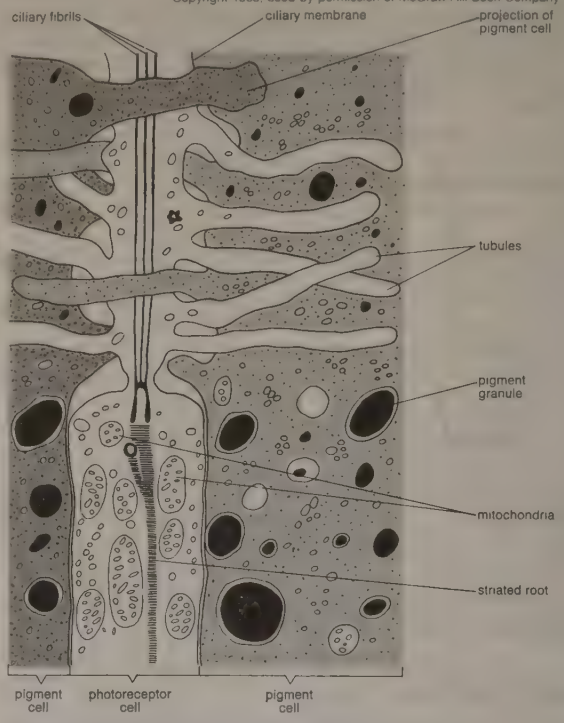


Figure 21: The distal part of a photoreceptor cell and two adjacent pigment cells in the ocellus of the hydromedusan *Polyorchis penicillatus*.

found in the cells of most animals; their motion is used either to move the animal or to move materials within or outside the body. As seen in the electron microscope, a cilium characteristically consists of nine double, peripherally arranged filaments (very fine threadlike structures) and a central pair of filaments in a so-called 9 + 2 pattern. A cilium arises from a cell structure known as a centriole, near which is often found a second centriole; ciliary rootlets extend from the centrioles to deep within the cell. In sensory cells a cilium often develops into a sensory organelle. Sensory cilia invariably lack the central pair of filaments, having what is called a 9 + 0 pattern. The sensory cilium of *Polyorchis* contains numerous coarse microvilli, which constitute the sensory portion; centrioles and a ciliary rootlet are also present. Near the base of the photoreceptors are frequently found large numbers of cell components called mitochondria. Mitochondria manufacture the compound adenosine triphosphate, which seems necessary for the photoreception process.

Photoreceptors in lower invertebrates

The microvilli found in the primitive photoreceptor of *Polyorchis* are atypical in two respects: first, it is unusual for the microvilli of rhabdomeres to be derived from cilia; second, the microvilli of the rhabdomere, which are usually packed together very densely, are, in the jellyfish, mingled with the microvilli of the pigment-containing cells that envelop the rhabdomere. Migration of pigment into or out of the pigment-cell microvilli provides a mechanism for regulating the amount of light absorbed by the receptor. The identification of photoreceptor cells in the jellyfish has been based on anatomical and behavioral evidence. The cells are located in ocelli, the removal of which renders the jellyfish incapable of responding to illumination.

Although the ctenophorans (comb jellies) usually have been assumed to lack both a response to light and photoreceptor cells, recent evidence has raised some doubt about such a conclusion. Electron microscopic examination of *Pleurobrachia pileus* has revealed a radial arrangement of four groups of lamellate bodies; the lamellae are composed of membranes of about 12 sensory cilia (*i.e.*, 9 + 0 type). The sensory cilia, instead of developing microvilli as in *Polyorchis*, become individual platelets. These lamellate structures, which are located in infolded regions of the presumed photoreceptor cell, resemble similar lamellate structures in certain molluscan and vertebrate photoreceptors; it is presumed by some investigators, therefore, that the lamellate structures of *Pleurobrachia* may be photoreceptors.

The planarian flatworm *Dendrocoelum lacteum* (phylum Platyhelminthes) has photoreceptor cells with well-developed rhabdomeres (see Figure 22). The cytoplasm of the receptor cell, which has a tubular extension, contains

From *Zeitschrift fur Zellforschung und mikroskopische Anatomie* (1961) in D. Mazia and A. Tyler (ed.), *General Physiology of Cell Specialization* (Copyright 1963), used by permission of McGraw-Hill Book Company

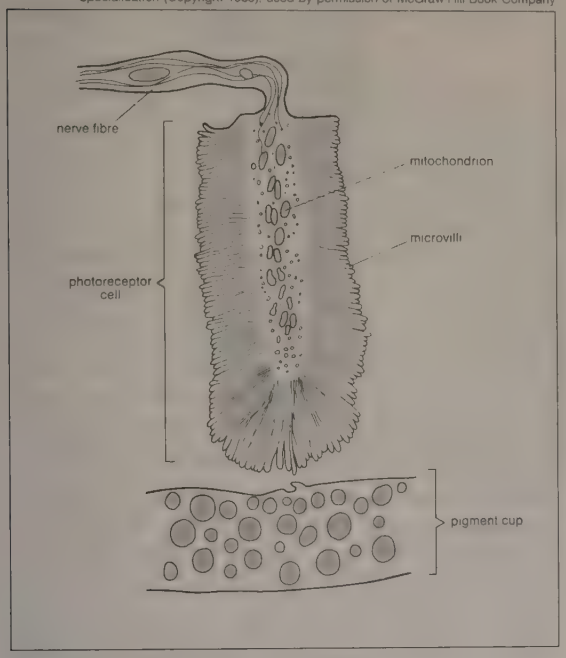


Figure 22: The photoreceptor cell in the flatworm *Dendrocoelum lacteum*.

many mitochondria; the surface membrane, which forms the tightly packed microvilli that constitute the photoreceptor organelle, is presumed to be the site of the visual pigment. No evidence has been found that either this photoreceptor or those of other flatworms are derived from cilia.

The nemertine worms (phylum Nemertea) have photoreceptor cells that resemble those of planarians when viewed with the light microscope. The photoreceptor cells on the eye of *Lineus ruber* resemble those of planarians in the electron microscope in that both have a rhabdomere consisting of microvilli. An important difference exists, however: the photoreceptors of *L. ruber* possess a filament that resembles a ciliary rootlet, even though no evidence exists that the rhabdomere is derived from cilia; in fact, cilia have not been observed in this cell. The function of the filament is not yet known.

Photoreceptors in the eye of the rotifer *Asplanchna brightwelli* (phylum Aschelminthes) are lamellate. Thin lamellae, arranged as leaves of a cabbage, apparently are folds of the receptor membrane. There is no evidence that the receptor is derived from cilia.

The photoreceptor cells in the eyes of the arrowworm *Sagitta scrippsae* (phylum Chaetognatha) have rhabdomeres consisting of structures of ciliary derivation that are called microtubules. The microtubules arise from a cone-shaped body filled with granules and cordlike structures. The cone-shaped body is derived from a sensory cilium. The microtubules comprising the rhabdomere are 50 nanometres in diameter and 20 long.

Photoreceptors of annelid worms

The segmented worms (Annelida) are the first invertebrate animal group in which the photoreceptors of a considerable number of species have been investigated. Studies indicate that annelids generally have rhabdomeres consisting of microvilli and are not of ciliary origin; however, at least one exception (*Branchiomma*) to this pattern has been observed.

The photoreceptor cell of the earthworm *Lumbricus terrestris* is an example of a rhabdomere comprised of microvilli; as mentioned previously, the microvilli of the membrane of the rhabdomere border on a vacuole within the receptor cell rather than on its outer surface. A number of sensory cilia are mixed with the microvilli and extend from the cell cytoplasm into the vacuole. The microvilli are not derived from cilia, and their function is as yet unknown. Although rhabdomeres consisting of microvilli have been found in several other annelids, no evidence of cilia has been reported. In the polychaete worm *Nereis vexillosa*, however, the microvillar rhabdomere arises from a tubular extension of the cell containing a centriole and a small fibre (fibril) with the structure of a ciliary rootlet; no direct evidence has been found thus far to link these ciliary vestiges to the rhabdomere.

The photoreceptor of the polychaete *Branchiomma vesiculosum* is a well-documented example of an annelid photoreceptor derived from cilia. The photoreceptor cell has a large invaginated (inpocketed) cavity filled with about 450 flattened lamellar sacs, which are the expanded flattened membranes of cilia having the 9 + 0 configuration. A collection of mitochondria and the nucleus of the receptor cell are displaced from the light path.

The Onychophora represent a transitional group that combines features of both the annelids and the arthropods. The photoreceptor cells of several species of onychophorans have been studied with the electron microscope. The photoreceptor cell of *Peripatus* closely resembles that of some annelids. The rhabdomere, which consists of microvilli, has a sensory cilium near its base. In studies of the *Peripatus* eye during development, no connection has been found between the sensory cilium and the developing microvilli. The function of the sensory cilium in this and in other photoreceptors in which no clear developmental relationship exists between the two is obscure.

Arthropods have few cilia, although they do occur in a few organs and provide the developmental basis for one type of arthropod ear, the chordotonal organ. Cilia have not been found in association with arthropod photoreceptors, however. The photoreceptor cells of all arthropods studied thus far are generally similar. Part of the cell surface forms densely packed microvilli, which is the photoreceptor containing the visual pigment.

The presumed photoreceptor in echinoderms is a collection of loosely packed microvilli extending from each receptor cell into the central cavity of the ocellus. In some species, each receptor cell has at least one cilium with the 9 + 0 pattern; this pattern is presumed to exist in other species. No relationship exists between the cilia and the microvilli; thus, the same condition prevails as in certain mollusks, annelids, and *Peripatus*. The echinoderms studied thus far appear to have a rhabdomere composed of microvilli and in close association with a sensory cilium.

Photoreceptor cells in mollusks and cephalopods

The mollusks (*e.g.*, scallops, clams, squid) have two morphologically distinct types of photoreceptor cells. One contains microvilli; of these, some photoreceptors have rudimentary cilia and fibrils such as those of certain annelids. Most mollusks appear to have this type of photoreceptor. The other type contains cilia; about 100 cilia on one surface of the cell develop into flattened sacs that are either packed tightly, like slices of bread, or curl on one another, like cabbage leaves. The ciliary type of rhabdomere is found in *Onchidium verruculatum* and in *Cardium edule*. The eyes of the bivalves *Pecten, Chlamys, Spondylus*, and *Amussium* contain double retinas, each with one of the types of photoreceptor. Evidence indicates that both types are photoreceptors and that they have different functions. The distal retina, with a ciliary

photoreceptor, responds to a decrease in illumination; the proximal retina, with a microvillar photoreceptor, responds to increasing illumination.

The cephalopod retina has long cylindrical photoreceptor cells with rhabdomeres consisting of microvilli. In some cells the long axes of the microvilli are horizontal. The other type has vertical microvilli. The axes of both types of microvilli are parallel to a plane that is tangent to the retinal surface. This orthogonal arrangement of the microvilli in different cells—*i.e.*, in which they are perpendicularly disposed to one another—forms the physical basis for polarized light detection in the cephalopods.

Vertebrate photoreceptors. The photoreceptor of such primitive chordates as the urochordate ascidian tadpole is a modified cilium with the 9 + 0 pattern; the cilium has developed into numerous lamellae. The cephalochordate *Branchiostoma californiense* may have both microvillar and ciliary photoreceptors. The Hesse cells, shown experimentally to be photoreceptor cells, have typical microvillar rhabdomeres with no cilia. Two other types of cells that may have a photoreceptor function are the dorsal ependymal cells (*i.e.*, special cells lining the cavities of the brain) and cells of the infundibulum, a ventral portion of the brain. Both cells have appendages containing cilia that form lamellae.

Other vertebrates have the two types of photoreceptor cells mentioned before: rods and cones. The rods are the photoreceptors for vision under conditions of dim illumination; the cones mediate daylight vision and colour sensation in many animals. The photoreceptors in both the rods and cones are composed of stacks of disks derived from cilia; the lamellae are stacked at right angles to the long axis of the cell and constitute the outer segments of rod and cone cells.

Most of the lamellae of the outer segment of the rod consist of free-floating disks within the limiting membrane of the outer segment (see Figure 23). The free-floating disks are formed at the base of the outer segment by

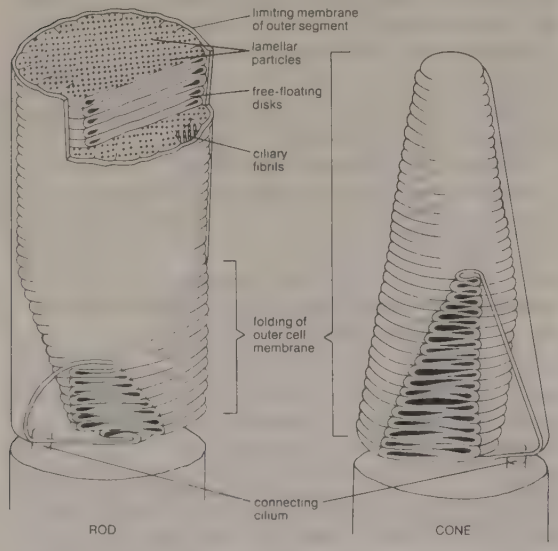


Figure 23: Structural relations between the outer segments of rods and cones (see text).

a process of continuous infolding of the membrane of the outer segment. The process of new disk generation proceeds continuously during the life of the animal, and the entire outer segment of the rod is replaced in about one week. As new disks are being formed at the base of the rod, disks at the other end of the outer segment are being broken off and ingested by pigment cells; the rod length thus remains constant. As infolded regions of the membrane at the base of the rod move outward, they are pinched off to form free-floating disks. The outer segments of the cones, sometimes conical, or cone-shaped, and shorter than the outer segments of the rods, do not have free-floating disks. The process of infolding stops just after

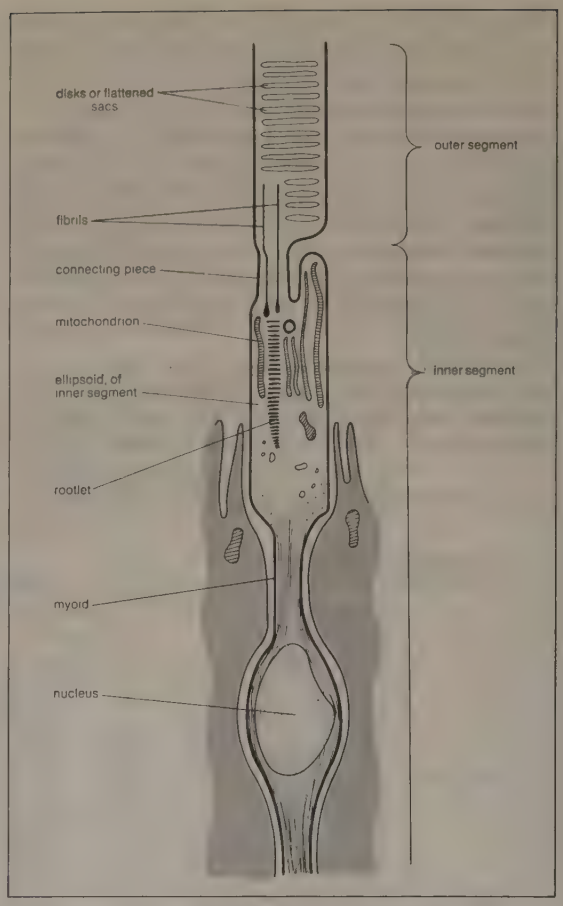


Figure 24: Part of a rod cell in the eye of a vertebrate.

From F S. Sjostrand in G K. Smelser (ed), *The Structure of the Eye*, Academic Press, Inc., in D Mazia and A Tyler (eds), *General Physiology of Cell Specialization*, Copyright 1963 McGraw-Hill Book Company, used by permission of McGraw-Hill Book Company

formation of the outer segments of the cones; infoldings occur along the entire length of the cones (see Figure 23). Because the membrane of the outer segments of the cones is not formed continuously during life, membrane components are continuously replaced. It has been shown that renewal takes place continuously along the entire surface of the membrane in cones. In rods, on the other hand, the renewal occurs only at the base, as new disks form.

Structural differences between rods and cones

The terms rod and cone are misleading, because these receptors cannot always be identified by shape; considerable variation exists between species. There are, nevertheless, a number of morphological differences between the rods and cones. The outer segments of the rods of many species are deeply incised—*i.e.*, lengthwise clefts in the disks give the outer segment a scalloped appearance in cross section. Both the rods and cones contain dense aggregations of mitochondria in a section of the inner segment called the ellipsoid (see Figure 24). In the same region the cones of many species have an oil droplet, sometimes coloured, that filters the light before it reaches the receptor. An important difference between the photoreceptors of vertebrates and those of most invertebrates is that the latter give rise to nerve fibres that synapse either with cells in the central nervous system or with cells in a nerve ganglion within the eye; the vertebrate receptors, on the other hand, end in a region called a foot piece, into which nerve cells of the retina send synaptic projections.

Embryology and evolution of photoreceptors. There are two types of vertebrate eyes. The more familiar is the highly developed lateral eye. The other is the primitive median, or pineal, eye (in the top of the head). Only cyclostomes, reptiles, and amphibians have a median eye. It is best developed in lizards, in which the cornea is translucent and the number of receptors is small. The photoreceptor cells of the pineal eye are homologous (*i.e.*, structurally or developmentally related) with those of the lateral eye. They are, in fact, cones, the outer segments

showing the characteristic infoldings along their entire length. In the median eye, the ends of the outer segments face the lens and light; in the lateral eye the retina is inverted, so that the ends of the outer segments face away from the light and toward the back of the eye.

The vertebrate retina is literally a part of the brain. The photoreceptor cells are derived from cells that line the neural tube, a hollow, dorsal structure that appears early in the embryo. The median eye develops from an outpocketing of the neural tube, and the cilia that develop into outer segments face the lens.

Invertebrates lack a neural tube; their photoreceptor cells develop from embryonic epithelial cells that send out axons—threadlike extensions—that grow into the central nervous system. Often the epithelial cells contain cilia, and the photoreceptors develop from the cilia. The ciliary photoreceptors of invertebrates do not appear to be homologous with those of vertebrates. As mentioned above, it is of particular interest that photoreceptors of diverse structure and in many different phyla develop either from cilia or in close association with them. Still unknown is the developmental and functional role of the sensory cilia that are frequently found in association with photoreceptors of apparently nonciliary origin.

Optical properties of photoreceptors. *Rhabdomeres and outer segments as light guides.* The densely packed membranous structures of photoreceptors have a higher refractive index than the surrounding substances. The refractive index of the rhabdomere of the blowfly (*Calliphora*), for example, is 1.349, and that for the surrounding substance is 1.336; any light reflected from the inside surface of the rhabdomere is called dense to rare reflection. When the angle at which the light strikes the inside surface (angle of incidence) is such that the change of the angle of the light when entering the photoreceptor (*i.e.*, the angle of refraction) is 90°, all of the light is reflected back into the rhabdomere. The angle of incidence at which total internal reflection occurs is called the critical angle. The rhabdomere and outer segment trap the light entering them at angles equal to or greater than the critical angle and propagate it, essentially without loss of energy. The rhabdomere and outer segment thus act as an optical wave guide, or light pipe (see OPTICS).

Although all light entering the photoreceptor organelle—*i.e.*, all light incident at angles equal to or less than the critical angle—is propagated by successive internal reflection, only certain angles of reflection occur; these are determined by various physical characteristics of the organelle. As a result, the light within the photoreceptor light guides is propagated in modes, or patterns, of energy. This is important for the photoreceptive function of the rhabdomere and outer segment because the amount of information the organelle can carry is related to the number of modes propagated. The thinnest photoreceptor organelles, such as those of certain fly rhabdomeres with a diameter of 0.5 nanometre, can, for example, support only one mode, called the lowest order mode. The amount of energy propagated in the rhabdomere in this mode depends on the wavelength of light, so that the shorter wavelengths are best propagated. Thus reddish light is not made available to the rhabdomere's visual pigment as readily as bluish light; the light guide properties of the organelle, therefore, can influence the physiological response of the photoreceptor. The light guide also profoundly affects the amount of image information transmitted because the greater the number of modes, the more image information. This last consideration is of particular importance for fused rhabdoms, in which image information apparently is not transmitted (although the question is not yet entirely settled). These light guide properties are of importance for still another reason. A small amount of the modal energy propagated by a light guide is actually outside the structure. Thus when, as described in the next section, pigment granules come near the photoreceptor organelle, they control the amount of light in the organelle by absorbing the energy propagating outside, so that it can no longer be transmitted by the photoreceptor. Such pigment granules may exert an even more profound effect by raising the refractive index outside the organelle, thus destroying its guiding properties and causing the light to spread into adjoining structures where it is absorbed, but not by visual pigment; absorption thus does not result in sensation.

Photomechanical light and dark adaptation. A mechanism that may partly control the amount of light within the photoreceptor under widely different illumination conditions involves the migration of photoprotective pigment granules either within the receptor or within neighbouring cells so as to envelop the photoreceptor; it has been observed in both vertebrates and invertebrates. Pigment migration in response to illumination takes place in the retinas of most vertebrates except mammals. The most pronounced effects are found in fishes, frogs, and toads; they also occur in reptiles and in diurnal and nocturnal birds.

Migration of pigment granules

Pigment migration is combined with photomechanical movements of the photoreceptor cells of many lower vertebrates (Figure 25). In darkness the pigment granules of the pigment epithelium are withdrawn, as shown in the figure. The rod myoids (contractile elements) are contracted in darkness so that the rod outer segments are withdrawn from the pigment and exposed fully to the available light. On the other hand, the cone myoids are extended, so that their outer segments are within the pigment, which functions effectively to reduce the amount of light in the outer segment, as explained in the previous section. The cone outer segments are enveloped by pigment, presumably to help suppress responses from the daylight receptors under conditions of dim illumination. Under the influence of light the rod myoids extend within several minutes, and the pigment envelops the rod myoids to help suppress the diurnal activity of the rod system (pigment surrounds only

From Dr. Samuel R. Detwiler *Vertebrate Photoreceptors* (1943), Macmillan & Co

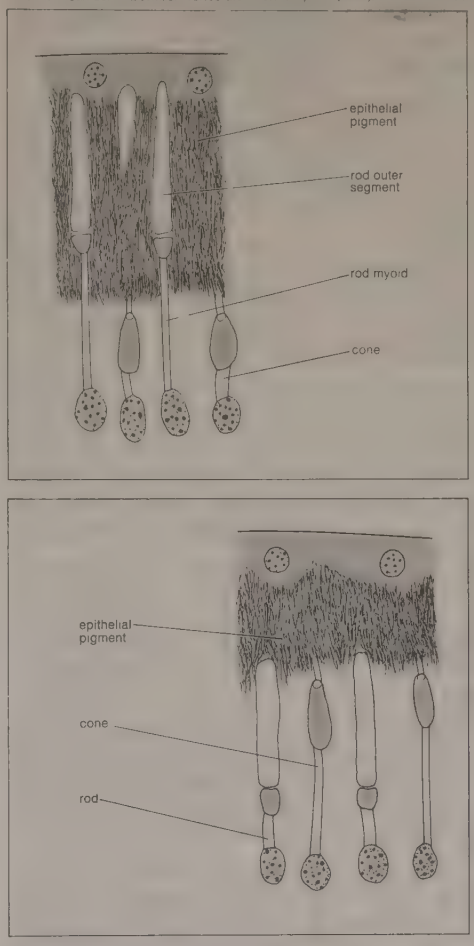


Figure 25: *Adaptation of retina.*
(Top) Theoretical light-adapted retina, showing migrated epithelial pigment, elongated rods, and contracted cones. (Bottom) Theoretical dark-adapted retina, showing contracted pigment, elongated cones, and contracted rods.

the rod myoids, not the outer segments [see Figure 25, top]). At the same time the myoids of the cones contract to withdraw the cone outer segments from the pigment. As mentioned above, neither photomechanical nor pigment movements occur in mammalian eyes; switching between rod and cone systems is accomplished neurally.

The envelopment of rhabdomeres by pigment migration is widespread in invertebrates. An example of this phenomenon occurs in the ommatidium of the amphipod *Gammarus ornatus*, in which the rhabdomeres of five radially arranged receptor cells form a star-shaped rhabdom. Extensions of the receptor cells are directed toward the cornea next to the lenslike crystalline cone and away from the cornea to cell bodies beneath a so-called basement membrane. In the dark-adapted condition the pigment is located around the cone and in the cell body. The rhabdom functions as a light guide within which is the visual pigment, which absorbs light. In the light-adapted condition, the pigment migrates from the receptor cell body to surround the rhabdom completely. Similar cell pigment migrations occur within the receptors of many invertebrates. In the compound eye of the horseshoe crab *Limulus*, the rhabdom is also star-shaped in cross section, but the rays of the star are much longer than are those of *Gammarus*. In the eye of *Limulus* the pigment migrates radially within the receptor cell, completely enveloping the rays in a matter of minutes in bright light and requiring about one hour to withdraw entirely from the rhabdom area in darkness.

The rhabdom of the migratory locust (*Schistocerca*) shows an interesting variation. In the light-adapted retina, mitochondria migrate and become tightly packed around the rhabdom. This may be a device for providing energy for the rapidly metabolizing light-activated photoreceptor. There is physiological evidence, however, that, in the light-adapted condition, the mitochondria, which have a high refractive index, increase the critical angle by coming close to the rhabdom. They thereby cause loss of light from the rhabdom and a decrease in the "field of view" of the rhabdom. When the eye is dark-adapted, the space around the rhabdom is replaced by fluid-filled spaces, with a low refractive index; they return the rhabdom to an efficient light guide mode of operation.

Birefringence and dichroism. The physical properties of certain crystals, glass, liquid, and gas, are the same regardless of the direction of the light propagated through them; for example, the speed and thus the refractive index of light are the same regardless of its direction of propagation. Such a medium is said to be isotropic. In other crystals, the atoms are arranged so the refractive index varies with the direction of propagation of light; such crystals are said to be anisotropic. If a medium has a crystalline arrangement that retards light in a particular direction, it is birefringent. In such an arrangement an entering ray of light is divided into two rays that are polarized in planes at right angles to one another.

When birefringence can be made to disappear by immersing a substance in a liquid with an appropriate refractive index, the birefringence is caused by an orderly arrangement of isotropic particles submicroscopic in size, such as the limiting membranes surrounding cells. Such birefringence is called form birefringence to distinguish it from the intrinsic molecular crystalline birefringence described above. Very small rods, lying parallel to each other, cause positive birefringence; parallel platelets cause negative birefringence. Studies have shown that rods in frogs have a negative form birefringence caused by the platelets and a positive intrinsic birefringence. These phenomena are thought to result from the ordered arrangement of a lipid (fat) layer two molecules thick in the platelet membranes.

Birefringence in the rods of frogs

Variation in the colour of light absorbed dependent on the direction of polarization of the light is termed dichroism. This property is a sensitive indicator of the orientation of molecules in a structure; dichroism in photoreceptors, for example, results from the ordered arrangement of the visual pigment molecules. The visual pigment of outer segments is dichroic, as are the outer segments. Dichroism of the outer segment can be demonstrated only by measuring the absorption of polarized light shone through the side of the outer segment. Light that propagates in the usual direction, that is down the long axis of the outer segment, does not show different absorption properties that depend on the direction of polarization. The fact that there is not dichroism for light propagating along the long axis, but there is for light propagating perpendicular to the long axis, indicates that the visual pigment molecules are oriented at random but with the long axes in the plane of the disks.

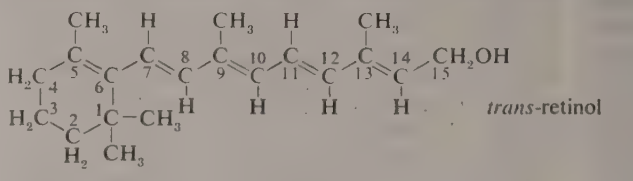
Rhabdoms consisting of microvilli can detect the plane of polarization of polarized light. Invertebrates with such rhabdoms use polarization properties of the blue sky opposite the Sun for navigation. In animals with the ability, the rhabdomeres comprising a rhabdom show orthogonal orientation. It has been shown that the microvilli are dichroic and that the dichroism is caused by the visual pigment. The orientation of the microvilli provides the mechanism for sensation of polarized light. Vertebrates cannot detect polarized light because of the random orientation of visual molecules within the plane of the platelets.

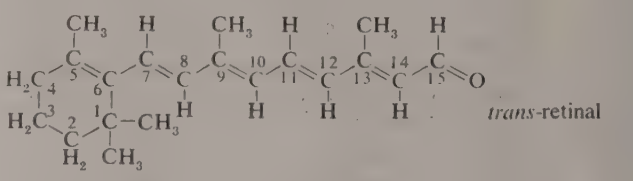
Stiles-Crawford effect. In the Stiles–Crawford effect, first observed in humans, the luminous efficiency of light has been shown to be greater when entering the centre than when entering the edge of the pupil. This effect may be considered an optical property of the cone receptor cells. The energy acceptance, or radiation, pattern of photoreceptors is a consequence of the optical wave guide properties resulting from the small size of the receptors. The Stiles–Crawford effect is not observed with dim illumination (rod vision), probably because the acceptance angle of the rods is wider, and it essentially fills the pupil.

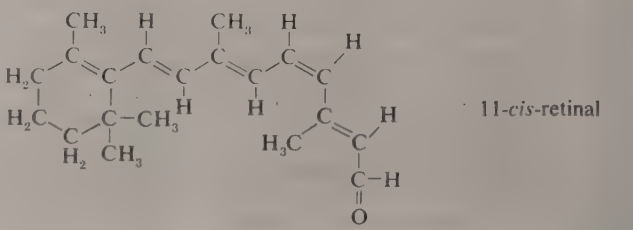
Visual pigments. As light propagates through the photoreceptor, the visual pigment absorbs photons (light particles). The absorption of one photon by a visual pigment molecule can initiate events that lead to nervous excitation. The photoreceptors in the eyes of both vertebrates and invertebrates consist of tightly packed membranes (lamellae), the function of which is to support the visual pigment and other molecules necessary for excitation. The photoreceptor membrane is a lipid layer two molecules thick, with protein either between the layers, coated over the layers, or both. The outer segment of the frog rod is 40 percent lipid and 60 percent protein; more than 80 percent of the protein is visual pigment.

Components. Visual pigment has two components: the light-absorbing chromophore (a chemical group that produces colour) and the protein moiety (opsin) to which it is chemically attached. The structure of the naturally occurring chromophore, retinal, is thought to be the same in all animals. Before it has been exposed to illumination, the chromophore is 11-*cis*-retinal, an alternate structure of vitamin A aldehyde. Upon exposure to light each

Characteristics of retinal


trans-retinol


trans-retinal


11-*cis*-retinal

chromophore molecule absorbs one photon of light. The absorption of a photon changes 11-*cis*-retinal to *trans*-retinal. (See structural formulas.) The resulting change in the shape of the molecule causes the separation of retinal from its protein component and gives rise to a number of coloured intermediates. In the dark, 11-*cis*-retinal combines spontaneously with the protein component to regenerate the visual pigment.

Hundreds of visual pigments have been identified and characterized by the wavelength of light that is best absorbed by them (absorption maximum, or λ max); the wavelength maxima range from 432 nanometres for the green rod of the frog to 625 nanometres for the red-absorbing cone of the goldfish. The reason for such a great range when the chromophore is the same in all visual pigments lies in part in properties of the chromophore; much of the variation, however, resides in the protein component. Variation in protein structure among different animals gives rise to different absorption maxima.

There are two forms of vitamin A: retinol A_1 and retinol A_2. The structure of A_1 differs only slightly from that of A_2. The effect of this slight structural difference is to increase the wavelength for maximum absorption by as much as 20 to 60 nanometres, depending on the pigment involved. The pigments with the shortest λ max are A_1 pigments; those with the longest are A_2. There is a broad area of overlap within the two groups, however. In general, marine fishes, land vertebrates, arthropods, and mollusks have A_1 pigments; freshwater fishes have A_2 pigments.

The protein components, or opsins, of visual pigments have not yet been fully characterized. The visual pigment of rods, a combination of retinal and opsin, is called rhodopsin, or visual purple. Estimates for the molecular weights of the opsins found in rods centre around 40,000. Squid opsin, however, has a molecular weight of about 70,000. Analyses of the visual pigments from the rods of squid and vertebrates suggest that the protein components are quite different. The nature of cone pigment proteins is thus far virtually unknown.

Important to an understanding of the photochemical reaction is the nature of the attachment of the chromophore to the opsin. Although controversy still exists on this point, it is clear that the visual pigment is part of the photoreceptor membrane and is very closely attached to its lipid and protein.

The role of visual pigments in vision. The sensitivity of the dark-adapted human eye can be measured as a function of wavelength of the stimulating light falling on that part of the retina in which the density of rods is greatest. When such a curve is compared with a curve of the absorption spectrum (the entire range of wavelengths of light absorbed) of human rhodopsin in the outer segments of the rods, the two curves correspond exactly. This is strong evidence that rhodopsin initiates the visual process for light absorbed in the outer segments.

The problem of cone pigments is an important one. More than 100 years have passed since it was concluded after a comparative study of retinas that the rods of vertebrate retinas mediate vision in dim light and the cones mediate daylight and colour vision. Recently, a method called microspectrophotometry has been developed for the study of cone pigments and other visual pigments. By projecting a minute beam of light onto a single photoreceptor, the absorption spectra of a variety of vertebrate and invertebrate photoreceptors have been measured. The absorption spectra of single human and other primate cones have thus been characterized. There are three types: a blue absorbing cone, a green absorbing one, and a yellow absorbing one. Originally suggested in 1802, these three cone types, now positively identified, can quantitatively explain human colour sensation.

Absorption spectra of photoreceptors

Such methods have been helpful in identifying receptors and their visual pigments in animals other than man; for example, in the retina of the frog *Rana pipiens*, a green rod and a red rod have been identified. A double cone with principal and accessory pigments and a single cone have also been characterized. Microspectrophotometry has been used to obtain evidence of visual pigments in rhabdomeres of the fly *Calliphora*.

PHYSIOLOGICAL RESPONSE OF PHOTORECEPTORS

The initial molecular events of photoreception can be summarized: after the absorption of a photon of light, 11-*cis*-retinal is converted into *trans*-retinal; the process disrupts the binding of retinal to opsin and results in the sequential formation of a number of coloured photoproducts.

The final physiological effects of photoreception have also been well defined. Excitation of visual pigments results in a change in certain properties (*i.e.*, permeability and potential) of the limiting membrane of the photoreceptor cell. Because the initial and final molecular events of photoreception are now understood, there is considerable hope that the entire chain of molecular events between the absorption of light by the visual pigment and the changes in potential of the cell membrane can eventually be reconstructed. This challenging problem has stimulated research concerned with the nature of the photoreceptor membrane and the chemistry of the photoproducts.

The internal environment of cells is generally different from that outside. The main cation (positively charged atom) inside cells is usually potassium; the main cation outside is sodium. In the resting, or unexcited, state, the cell membrane of most excitable cells (such as receptor, nerve, and muscle cells) is impermeable to sodium ions but permeable to potassium ions. A separation of charges occurs on either side of the membrane because the positively charged potassium ions diffuse outside of the cell, leaving a surplus of nondiffusing negatively charged atoms (anions) inside the cell. The result is a difference in potential across the cell membrane; this difference is proportional to the logarithm of the ratio of the concentration of potassium ions outside to that inside the cell. Photoreceptor cells are negative by some 30 to 60 millivolts (1 mV = 0.001 volt) inside relative to the outside, the exact value depending on the specific cell. This negative potential difference is termed the resting potential.

Receptor potential. The absorption of a single photon by the visual pigment in a photoreceptor may result in a physiological response of the receptor cell. Such a response of the cell can be measured as the receptor potential—that is, a change in the potential of the cell membrane caused by a change in the permeability of some part of the membrane to certain ions. The permeability change is in turn related to the absorption of light by the chromophore through a chain of molecular events as yet unknown. The receptor potentials of the eyes of animals from a number of phyla have been investigated. In certain photoreceptors the receptor potential is positive in sign, or depolarizing; in others it is negative, or hyperpolarizing.

Definition of receptor potential

Depolarizing receptor potentials. The depolarizing receptor potential, a change in potential response to illumination that is positive in sign, is characteristic of photoreceptor cells in many invertebrate phyla. One of the best known photoreceptors is the retinular cell in the compound eye and in the ventral eye of the horseshoe crab *Limulus*, an arthropod. The membrane potential of this cell is easily measured by means of a very small electrode (a device that can be placed inside a cell to measure the potential across the cell membrane). When the electrode penetrates the cell in the absence of light, the potential falls to some negative value, perhaps −40 millivolts, which is the resting potential. Even in complete darkness, however, variation in the resting potential occurs in the form of slow fluctuations. Although these spontaneously and randomly occurring fluctuations presumably seldom result in a physiological visual response, they are miniature receptor potentials. Each is accompanied by a decrease in the resistance of the membrane (*i.e.*, it becomes a better conductor) and an increase in the permeability of the membrane to sodium ions. When the receptor is illuminated with dim light, the slow potential fluctuations increase in frequency, and, with brighter light, they become briefer and higher in frequency, finally fusing to form a more or less steady depolarizing receptor potential. Thus, a short time after the cell is exposed to light, the membrane becomes depolarized; when the light is removed, the receptor potential decays, and the resting potential is restored. The extent of the receptor potential increases with increasingly stronger illumination; it has a linear relationship to the logarithm

Relationship between receptor potential and illumination

of the intensity of illumination. Whatever properties are responsible for this transformation are responsible for the photoreceptor's enormous dynamic range: despite the fact that the receptor potential can change at most on the order of 100 millivolts, the receptor can respond over a range of more than 10 orders of magnitude, a range of intensities of over 100,000,000,000.

In all species with this depolarizing receptor potential, the fact that the resistance of the photoreceptor cell membrane decreases during the receptor potential suggests that it becomes more permeable to some ion or ions during this time. Evidence as to the nature of the permeability change suggests that it is principally an increase in permeability to sodium ions that causes the receptor potential. There is additional evidence for this conclusion (see *Photoreception* in *Bibliography*).

The depolarizing receptor potential is characteristic of the arthropod photoreceptor consisting of microvilli and the receptor cell. Species from most of the orders of Arthropoda have now been investigated, and exceptions have not been reported. Usually, the photoreceptor cell gives rise only to a receptor potential. The receptor potential is conducted passively a short distance to higher order nerve cells via a nerve fibre. The higher order nerve cells convert the receptor potential to propagated action potentials (nerve impulses) for transmission of the information over longer distances.

The proximal sense cell of the double retina of *Pecten* is similar to the above in that illumination gives rise to a depolarizing receptor potential and a decrease in resistance. The morphology of the proximal sense cell is also similar to that of the arthropod photoreceptors in that it has coarse microvilli. But the proximal sense cell of *Pecten* has a long nerve fibre that carries nerve impulses. The depolarizing receptor potential activates the mechanism that initiates an action potential, and steady depolarization of the axon gives rise to a series of nerve impulses; the stronger the intensity of illumination, the greater the extent of the receptor potential and the higher the frequency of nerve impulses. Nerve activity of this nature is rarely found in arthropod photoreceptor cells.

In addition to the arthropods and *Pecten*, a proximal sense cell in the leech (*Hirudo*), an annelid, has the depolarizing type of receptor potential.

Hyperpolarizing receptor potential with decrease in resistance. As noted above, the proximal sense cells of the *Pecten* eye have a depolarizing receptor potential, and the nerve responds to illumination with a series of nerve impulses. The distal sense cells, which also have nerve fibres, respond to a decrease in illumination with nerve impulses. This is called an off response. Recent studies indicate that the receptor potential of the distal sense cell is a hyperpolarizing potential, a potential change that is negative in sign. When the cell is depolarized, the extent of the receptor potential increases; when the cell is sufficiently hyperpolarized, the receptor potential reverses polarity and becomes depolarizing. The receptor potential of the distal sense cell is accompanied by a decrease in resistance. The hyperpolarizing potential is believed to be caused by an increase in permeability to potassium and perhaps to other ions.

The off response

The photoreceptor of the *Pecten* distal sense cell is a ciliary derivative. A primitive chordate, the tunicate *Salpa*, which has a microvillar photoreceptor, has the same type of hyperpolarizing photoreceptor potential. There thus does not seem to be any association of physiological potential types with structurally or embryologically similar photoreceptors.

Hyperpolarizing receptor potential with increase in resistance. The rod and cone photoreceptors of vertebrate eyes have a third type of receptor potential. In the dark, sodium ions flow steadily into the outer segment; light decreases this dark current—*i.e.*, the cell becomes hyperpolarized, and its resistance increases. If the cell is sufficiently hyperpolarized under experimental conditions, depolarization occurs; the receptor potential thus is reversed, as would be expected if the permeability change involves principally sodium ions.

Interpretation of the physiological response. As a nerve cell is depolarized, it discharges nerve impulses; as it becomes hyperpolarized, the membrane potential returns to the resting potential. The fact that certain photoreceptor cells are hyperpolarized by illumination is thus confusing and raises the possibility that light is inhibiting the receptors. In the generation and transmission of information about illumination, however, the sign of the response may actually be unimportant. It has been known for some time that there are several types of optic nerve fibres in vertebrate eyes; some discharge impulses during illumination, others discharge either only at the onset and cessation of illumination or merely at the cessation. In other words, information about the cessation of illumination is at least equal in importance to information about its onset and duration. Such observations of optic nerve activity prove that the hyperpolarizing response of the receptor, the only response of the receptor, is translated into a code that includes the discharge of nerve impulses both during and after illumination. Thus, the hyperpolarizing response of the receptor, although it may be important in emphasizing the off response, does not result in the loss of information about excitation. Because the hyperpolarizing response is definitely used to produce excitation at a later stage, it is misleading to consider it an inhibitory response. The same is true in reverse for the depolarizing response.

The way in which initial photoreception is processed by the retina and higher centres to cause visual perception is beyond the scope of this article. It may be mentioned, however, that the initial information used in this processing is the direct result of the physical and chemical properties of the photoreceptor. (W.H.M.)

Sound reception

Sound waves are a particular kind of mechanical activity consisting of vibrations in a gas, liquid, or solid medium. If an animal possessing an auditory mechanism comes in suitable contact with a medium vibrating at a frequency and intensity within its range of aural (hearing) sensitivity, it may hear the sound. For land animals, the usual vibrating medium is the air; for fishes and other aquatic creatures, it commonly is the water. Yet, under suitable conditions, all hearing animals can perceive sound waves transmitted by media other than the one in which they live; thus, humans can hear noise while underwater. (For a detailed discussion of human sound reception, see below *Human hearing*; additional information is contained in the article SOUND.)

In the course of evolution, animals have developed a variety of sense organs that respond to mechanical stimuli. There are at least 10 of these mechanoreceptors in vertebrates and perhaps as many in advanced invertebrates. Not all of these structures respond to sound, however, for among them are the simple touch endings of the skin and the motion receptors that serve (mediate) bodily equilibrium (see above *Mechanoreception*). Although the different ways of registering mechanical changes in the environment or within the body represent various structural specializations, it is not feasible to identify any one of them simply in terms of its structure; many different mechanisms, cells, or organs may perform similar functions. Ears, for example, take many forms in the lower animals and often have little resemblance to these organs in man and other higher vertebrates. Yet the service that they perform in sound reception is similar enough that they may be called ears.

Although there is no fossil record of the origin and development of auditory structures, in animals with ears the evolutionary process in every instance appears to have been a conversion to an auditory function of structures that previously mediated a simpler form of mechanoreception. Indeed, any mechanoreceptor, even though best adapted to respond to some other form of mechanical stimulation, will respond to vibrations within some region of the sound frequency range if the vibrations have a sufficiently high level of intensity.

Many attempts have been made to define hearing, often with indifferent success. The task is difficult, and in certain respects the lines of distinction are arbitrary. The

Attempts to define hearing

ear cannot be identified by any standard structure, nor can it be identified in terms of the stimulus as simply a receiver of sound vibrations. As noted above, mechanical receptor organs will respond to sound vibrations within some region of the frequency range if a sufficiently high level of intensity is provided. Moreover, the ear cannot be characterized in terms of the physical principles by which it operates because these principles vary among the ears of different animal species.

A definition of hearing, therefore, must be sought in terms of the ear's specialization of function and the relative effectiveness with which it performs this function. Thus, hearing may be characterized as the reception of sound vibrations by an organ, the ear, that has developed for this particular purpose and that has reception of sound as its primary function. This definition excludes the reception of sound vibrations by touch (tactual) endings in the skin, for example, because these structures respond most readily to direct pressure. Before such receptors will respond to sound waves, the vibrational intensity of the sound must be relatively great. Also excluded are the hair sensilla, of which arthropods have many types, whenever it can be shown that these organs respond with greater sensitivity to another stimulus (most often a simple direct deflection of the central hair).

Principles of sound reception

Theoretically, several aspects of vibration might serve in its detection by an ear. These characteristics include the amplitude (extent) of the motion of particles (*e.g.*, molecules) in a medium, the velocity and acceleration of the motion, the pressure exerted upon an obstacle in the path of the sound waves, and temperature changes occasioned by the vibrations. All of these manifestations have been utilized in attempts to design microphones for the detection and measurement of sound, but only two (pressure and velocity effects) have proved to be of any practical value. Thus, those devices that employ these two effects are known as pressure and velocity microphones.

It seems more than coincidence that these same two aspects of sound, pressure and velocity, are the only stimulus characteristics on which the evolution of ears appears to have been based. Moreover, just as the pressure microphone is the most practical type designed by man, among ears the pressure type is the most widespread and the most highly developed. Ears that distinguish changes in velocity have appeared only in a few lower animals—as an elaborated hair organ in some insects and perhaps spiders and in two special forms among fishes. All other ears are pressure receptors that have taken two lines of evolutionary development, one in most of the insects and another in vertebrates above fishes.

Types of animals that have ears. Considering the usefulness of the sense of hearing to such highly organized animals as man, it may seem surprising that this sense is so limited in its appearance and development among animals. It is found only in two major groups of animals: arthropods (*e.g.*, insects and crabs) and vertebrates (*e.g.*, amphibians, birds, and mammals). The condition that probably limited the development of hearing in other species was the lack of sufficient advancement and flexibility of the nervous system.

Biological value of ears

In those animals with auditory structures, hearing serves purposes of great biological value: in its more primitive forms, it is used to sense danger and enemies, to detect prey, and to identify prospective mates; at a more complex level, hearing is involved in communication within social groups and in emotional expressions of various kinds. The cry of an infant mouse that has strayed from the nest elicits a response by the mother to retrieve it. The singing of a male thrush asserts a claim to its territory, attracts a female to the area, and warns off other males. Among higher mammals (*e.g.*, monkeys and apes) vocalizations show even greater variety and express a range of meanings that may be interpreted in human terms as expressions of such concepts as danger, aggression, love, and the availability of food. In man, the elaborations of auditory communication can be even more symbolically complex, extending to speech and music (see below *Human hearing*; see also SPEECH). The significant features in complicated sounds that people perceive and differentiate correspond to the physical dimensions of frequency (the number of waves, cycles, or vibrations per second), intensity, phase, complexity of wave form, and temporal pattern. The variety of distinguishable acoustic forms is enormous.

Among the most highly refined applications of the auditory sense are those found in such animals as bats and dolphins. These creatures are able to discern objects around them by a process called echolocation: the animal sends out a cry and, by the nature of the echo, is informed of the presence of obstacles or potential prey. For these animals, the sense of hearing provides a service in the dark that closely approaches the reliability of vision in the perception of objects and spatial relations.

ORGANS OF SOUND RECEPTION IN INVERTEBRATES

It has long been believed that at least some insects can hear. Chief attention has been given to those that make distinctive sounds (*e.g.*, katydids, crickets, and cicadas) because it was naturally assumed that these insects produce signals for communication purposes. Organs suitable for hearing have been found in insects at various locations on the thorax and abdomen and, in one group (mosquitoes), on the head.

Among the many orders of insects, hearing is known to exist in only a few: Orthoptera (crickets, grasshoppers, katydids), Homoptera (cicadas), Heteroptera (bugs), Lepidoptera (butterflies and moths), and Diptera (flies). In the Orthoptera, ears are present, and the ability to perceive sounds has been well established. The ears of katydids and crickets are found on the first walking legs; those of grasshoppers are on the first segment of the abdomen. Cicadas are noted for the intensity of sound produced by some species and for the elaborate development of the ears, which are located on the first segment of the abdomen. The waterboatman, a heteropteran, is a small aquatic insect with an ear on the first segment of the thorax. Moths have simple ears that are located in certain species on the posterior part of the thorax and in others on the first segment of the abdomen. Among the Diptera, only mosquitoes are known to possess ears; they are located on the head as a part of the antennae.

Location of insect ears

All the insects just mentioned have a pair of organs for which there is good evidence of auditory function. Other structures of simpler form that often have been considered to be sound receptors occur widely within these insect groups as well as in others. There is strong evidence that some kind of hearing exists in two other insect orders: the Coleoptera (beetles) and the Hymenoptera (ants, bees, and wasps). In these orders, however, receptive organs have not yet been positively identified.

Types of insect auditory structures. Four structures found in insects have been considered as possibly serving an auditory function: hair sensilla, antennae, cercal organs, and tympanal organs.

Hair sensilla. Many specialized structures on the bodies of insects seem to have a sensory function. Among these are hair sensilla, each of which consists of a hair with a base portion containing a nerve supply. Because the hairs have been seen to vibrate in response to tones of certain frequencies, it has been suggested that they are sound receptors. It seems more likely, however, that the sensilla primarily mediate the sense of touch and that their response to sound waves is only incidental to that function.

Antennae and antennal organs. Many sensory functions have been attributed to the antennae of insects, and it is believed that they serve both as tactual and as smell receptors (see above *Chemoreception*). In some species, the development of elaborate antennal plumes and brushlike terminations has led to the suggestion that they also serve for hearing. This suggestion is supported by positive evidence only in the case of the mosquito, especially the male, in which the base of the antenna is an expanded sac containing a large number of sensory units known as scolophores. These structures, found in many places in the bodies of insects, commonly occur across joints or body segments, where they probably serve as mechanoreceptors for movement. When the scolophores are associated with any structure that is set in motion by sound, however, the arrangement is that of a sound receptor.

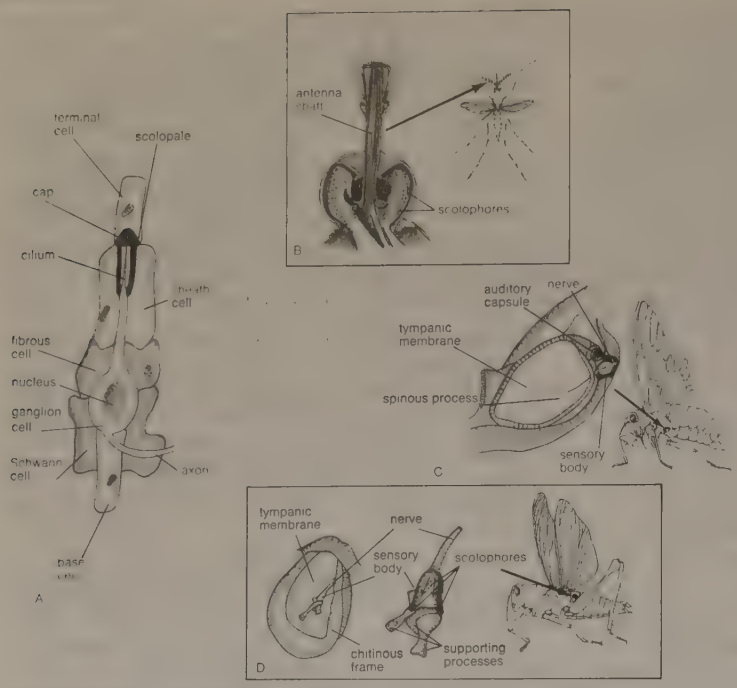


Figure 26: *Auditory mechanisms in insects.*
(A) A scolophore organ. (B) The mosquito ear. (C) The ear of the cicada *Magicicada septendecim*. (D) The ear of the grasshopper.

By courtesy of (B, right) the U.S. Department of Agriculture, and (B, left) *Acustica*, (1953) pages 335–343, and (C, right) H Weber, *Grundriss der Insektenkunde* (4th ed. Ed. Gustav Fischer, Stuttgart,1966), and (D, right) reprinted from Anna Botsford Comstock *Handbook of Nature Study*, Copyright, 1939, by Comstock Publishing Company, Inc. Used by permission of Cornell University Press, and (D, left and centre) E.G. Gray, *Philosophical Transactions* B 243, 77 (1960)

Structure of the scolophore

The basic structure of the scolophore is shown in Figure 26. Four cells (base cell, ganglion cell, sheath cell, and terminal cell), together with an extracellular body called a cap, constitute a chain. Extending outward from the ganglion cell is the cilium, a hairlike projection that, because of its position, acts as a trigger in response to any relative motion between the two ends of the chain. The sheath cell with its scolopale provides support and protection for the delicate cilium. Two types of enclosing cells (fibrous cells and cells of Schwann) surround the ganglion and sheath cells. The ganglion cell has both a sensory and a neural function; it sends forth its own fibre (axon) that connects to the central nervous system.

In the mosquito ear (Figure 26) the scolophores are connected to the antenna and are stimulated by vibrations of the antennal shaft. Because the shaft vibrates in response to the oscillating air particles, this ear is of the velocity type. It is supposed that stimulation is greatest when the antenna is pointed toward the sound source, thereby enabling the insect to determine the direction of sounds. The male mosquito, sensitive only to the vibration frequencies of the hum made by the wings of the female in his own species, flies in the direction of the sound and finds the female for mating. For the male yellow fever mosquito, the most effective (*i.e.*, apparently best heard) frequency has been found to be 384 hertz, or cycles per second, which is in the middle of the frequency range of the hum of females of this species. The antennae of insects other than the mosquito and its relatives probably do not serve a true auditory function.

Cercal organs. The cercal organ, which is found at the posterior end of the abdomen in such insects as cockroaches and crickets, consists of a thick brush of several hundred fine hairs. When an electrode is placed on the nerve trunk of the organ, which has a rich nerve supply, a discharge of impulses can be detected when the brush is exposed to sound. Sensitivity extends over a fairly wide range of vibration frequencies, from below 100 to perhaps as high as 3,000 hertz. As observed in the cockroach, the responses to sound waves up to 400 hertz have the same frequency as that of the stimulus. Although the cercal organ is reported to be extremely sensitive, precise measurements remain to be carried out. It is possible, nevertheless, that this structure, which is another example of a velocity type of sound receptor, is primarily auditory in function.

Tympanal organs. The tympanal organ of insects consists of a group of scolophores associated with a thin, horny (chitinous) membrane at the surface of the body, one on each side. Usually the scolophores are attached at one end by a spinous process to the tympanic membrane (eardrum); the other ends rest on an immobile part of the body structure. When the membrane moves back and forth in response to the alternating pressures of sound waves, the nerve fibre from the ganglion cell of the scolophore transmits impulses to the central nervous system. Because the tympanic membrane is activated by the pressure of sound waves, this is a pressure type of ear.

Simple tympanal organs, such as those found in moths, contain only two or four elements, or scolophores. In cicadas, on the other hand, these organs are highly developed; they include a sensory body (a number of scolophores in a capsule) that may contain as many as 1,500 elements. Shown in Figure 26 is the structure of the tympanal organ in the 17-year cicada (*Magicicada septendecim*).

With 80 to 100 scolophores, the grasshopper ear, which has been studied more thoroughly than any other insect ear, is structurally between that of moths and cicadas. Ordinarily, the tympanic membrane is hidden beneath the base of the insect's wing cover. A bundle of auditory nerve fibres runs from one side of the sensory body, which lies on the inner surface of the membrane (Figure 26), and joins other nerve fibres of the region to form a large nerve extending to a ganglion (nerve centre) in the thorax.

Evidence of hearing and communication in insects. *Behavioral observations.* That the insect ear serves an auditory purpose has been proved by a large number of experimental observations, particularly those that have dealt most extensively with katydids and crickets. Males of these groups produce sounds by stridulation, which usually involves rubbing the covers of the wings together in a particular way. One wing has a serrated surface (a "file") that runs along an enlarged vein; the other wing has a sharp edge over which the file is scraped. The scraping causes the wing surfaces to vibrate; the natural resonances of the vibrations and the particular rhythm and repetition rate of the scraping movements determine the nature of the song, which varies with each species. Most females

Stridulation

are silent, but those of a few species have a poorly developed stridulatory apparatus, and weak sounds have been reported. Both males and females have tympanal organs for sound reception.

The observation that the males of many insect species produce repeated stridulatory sounds during the mating season led to the inference that the primary purpose of these noises was to attract a female. That this is indeed the case was first established by the extensive observations of the Yugoslavian entomologist Ivan Regen, who worked over the period 1902–30 mostly with a few species of katydids and crickets. In one of his earliest experiments, Regen proved (1913–14) that a male katydid of the species *Thamnotrizon apterus* responds to the sound of another male by chirping. The first male responds in turn to the second male's chirp, and the two insects then set up an alternating pattern of chirping. Although this pattern had been observed earlier, Regen was the first to prove by a series of experiments that it depends upon the sense of hearing. After removing the forelegs, on which the tympanal organs are located, of certain males, he found that even though these insects continued to stridulate, they did so only in individual rhythms that were not affected by the sounds of other males. Any alternation of chirping between deafened males, or between a deafened and a normal male, occurred only rarely, for brief times, and by chance.

A long series of check experiments by Regen showed that other stimuli, such as light, odours, and surface vibrations, did not affect the chirping behaviour. In these experiments the insects were placed in separate rooms, and their sounds were transmitted by telephone.

Chirping and mating behaviour

Further experiments carried out by Regen on field crickets (*Liogryllus campestris*) demonstrated the reactions of females to chirping males. In the most elaborate of these experiments, 1,600 sexually receptive females were released around the periphery of a large enclosed area in the middle of which had been placed a cage containing one or more chirping males. Precise data concerning the frequency with which the females moved toward the cage were obtained by surrounding the cage site with an array of traps in which the females were caught as they moved inward. The results were statistically significant. Normal females (those with intact tympanal organs) moved toward the cage and eventually reached it. The removal of one foreleg and its tympanal organ, however, caused difficulty; the movements were more random and the approaches fewer, although some females did succeed in reaching the cage. When both tympanal organs were removed or if the male failed to chirp, the performance of the females was reduced to chance. They also failed to exhibit the seeking performances if the male's stridulatory organ was modified, as by removing the file, so that little or no sound was produced.

In 1926 Regen returned to his study of the alternating chirping pattern of katydids and succeeded in having males react to an artificial sound, one that Regen himself produced. He also found that the alternation could be demonstrated with a suitably active male by using a variety of sounds—whistles, percussion noises, and sounds made with his mouth. It was never altogether clear, however, what changes Regen had made in his signals that finally brought success; probably the secret lay in the particular rhythm and timing of the signals. At any rate, this method made possible a study of the general nature of the auditory sensitivity of these insects and the range of sound frequencies to which they responded. It was shown that katydids are most sensitive to the very high frequencies, those that are beyond the limit of the human ear. The instruments available to Regen at the time, however, did not permit a precise measurement of intensity thresholds. (A threshold is the lowest point at which a particular stimulus will cause a response in an organism.)

Although the work of Regen and others established the basic character of sound reception in insects and its role in communication and mating, other details had to await the introduction of electrophysiological methods in this field as well as the development of electronic methods for the precise production, control, and measurement of sound stimuli.

Electrophysiological observations. When making electrophysiological observations of an auditory mechanism, an electrode (one terminal, generally a fine wire, in an electric circuit) is placed on a nerve or some other sensory structure in the mechanism. Sounds, presented at different frequencies and intensities, produce neural or sensory changes, which are actually electrical discharges or changes in electrical potential of extremely small magnitude. The impulses are picked up by the electrode and transmitted to an instrument with which they can be amplified, observed, and recorded. In both behavioral and electrophysiological observations, the auditory sensitivity of an animal to sounds of different frequencies can be illustrated by a curve.

The electrophysiological method was first used in research on the insect ear in 1933, with observations mainly on two katydid and one cricket species. The tympanal organ of these insects is located on one of the segments of the foreleg; its nerve goes to a ganglion in the thorax. When an electrode is placed on this nerve, its threshold sensitivity and overall frequency range can be determined by varying the intensity and frequency of the sounds applied to the tympanic membrane. It has been found that the tympanal organ of these insects responds poorly to low tones (those of low frequency) but improves rapidly as the frequency increases to a maximum sensitivity around 3,000 to 5,000 hertz. For higher frequencies the sensitivity declines, until a limit is reached at 30,000 hertz. It is likely that the insect's identification of its own species by means of song is primarily in terms of intensity and time patterns, with the rapid changes of intensity playing a prominent part. The possibility of frequency also entering into the pattern, however, cannot be ruled out.

A further question concerns the perception of the direction of a sound source. Clearly, if a female is to seek out and find a chirping male, the effectiveness of her performance depends upon an ability to localize the sound. Experiments indicate that the magnitude of electric responses from the tympanal nerve in katydids varies in a systematic manner when a given sound is presented at different angles while the distance is held constant. The insects continue to exhibit this directional pattern even after one of the tympanal organs has been removed. As was mentioned earlier, Regen found that female crickets deprived of one tympanal organ were still able to locate a chirping male, though less effectively than when both organs were intact.

Evidence of hearing and communication in spiders. Whether spiders have a sense of hearing has long been debated. Early anecdotal observations concerning this matter have now been reinforced with both behavioral and electrophysiological evidence showing without doubt that spiders are sensitive to mechanical vibrations and also to aerial sounds. Whether this sensitivity should be regarded as hearing is considered later in this section, after a review of the anatomical and behavioral evidence.

Anatomical evidence. The bodies of spiders contain many slitlike openings, called lyriform organs, that have been considered as sensory in nature. Most of these organs probably have a kinesthetic function and thus provide information on local movements of body parts. There is one type of lyriform organ, however, that differs from the others in its location and in certain structural details. It is found on the metatarsal (next to last) segment of each of the eight legs, close to the joint that this segment makes with the tarsus (the last segment, or foot), and consists of a number of slits—about 10 in the common house spider—that partially encircle the leg. Each slit contains a fluid chamber the inner wall of which is pierced by a tubule through which a thin filament runs to one of the two side walls (lamellae) that enclose the slit. This filament is evidently the termination of a ganglion cell that lies deeper in the leg. It has been suggested that an alternating compression of the lamellae stimulates the terminal filament.

Lyriform organs

The responsiveness of the common house spider to aerial sounds and mechanical vibrations includes a wide range, from below 20 to as high as 45,000 hertz. Within this range the sensitivity, as measured by electrical potentials, varies widely for aerial sounds; in some experiments

narrow regions of frequency have been found in which no responses could be obtained at the highest intensities available. These variations of sensitivity are ascribed to mechanical resonances in the lyriform structure.

The tarsus evidently plays an important part in responses to sounds. Removal of portions of the tarsus reduces the responses about in proportion to the amount removed; immobilization of the tarsus greatly impairs the sensitivity. It appears, therefore, that the tarsus serves as a sensing element that transmits vibrations to the lyriform organ, which thus is a velocity type of ear.

Behavioral evidence. It has been reported that spiders react in characteristic ways to a buzzing insect caught in their web. The spider apparently locates the insect at once, runs to it, and attacks it. An inactive object, however, such as a small pebble enmeshed in the web, produces a different response: the spider manipulates the strands of the web, locates the object, and cuts away the filaments surrounding it so that the object drops to the ground. The reactions of a house spider to a mechanical vibrator applied to a point on the web have been observed. Such a stimulus elicits a response similar to that of an active insect if the vibratory frequency is between 400 and 700 hertz. For frequencies above 1,000 hertz, however, the spider reacts either by running to a secluded corner of the web or, if the intensity is too great, by abandoning the web altogether. From this and similar evidence it has been concluded that the spider has the ability of pitch (tone) discrimination between low and high ranges and perhaps can distinguish between tones of the lower range.

Spiders also react to aerial tones from an artificial source, such as a loudspeaker. These stimuli elicit an orientation response, in which the spider faces the source and reaches out with the two front legs. Thus, in view of the high level of sensitivity to both aerial and mechanical stimuli, the reception of sounds in the spider can probably be regarded as true hearing, and the lyriform organ as a form of ear. It is evidently a velocity type of ear, for there is no tympanic surface to respond to sound pressures, and the small leg segments seem to respond to the oscillatory motions of the air particles.

SOUND RECEPTION IN VERTEBRATES—AUDITORY MECHANISMS OF FISHES AND AMPHIBIANS

The labyrinth

The ear of vertebrates appears to have followed more than one line of evolutionary development, but always from the same basic type of mechanoreceptor, the labyrinth. All vertebrates have two labyrinths that lie deep in the side of the head, adjacent to the brain. They contain a number of sensory endings the primary functions of which are to regulate muscle tonus (a state of partial muscular contraction) and to determine the position and movements of the head and body.

Generalized sketches of vertebrate labyrinths are shown in Figure 27, with the usual locations of the sensory endings indicated for the different vertebrate classes. Two main divisions of these endings are distinguished: a superior division, which includes the three semicircular canals, the organs associated with the sense of balance, and the utricle, a small sac into which the semicircular canals open; and an inferior division, which includes the saccule (also a small sac) and its derivatives. Arising at or near the connection between the utricle and the saccule is the endolymphatic duct, which ends in an endolymphatic sac; this structure probably regulates fluid pressures in the labyrinth and aids in the disposal of waste materials.

The superior division of the labyrinth (Figure 27) is remarkably constant in form throughout the vertebrates except in the cyclostomes (*e.g.*, hagfishes and lampreys), in which the canals and endings are reduced in number. The utricle contains a macular ending, the macula utriculi, and each semicircular canal ends in a crista. In all vertebrate classes except the placental mammals and a few other scattered species, a papilla neglecta is present. It is usually located on the floor of the utricle or near the junction of the utricle and the saccule.

The inferior division of the labyrinth always contains a saccule with its macula, the macula sacculi, but the derivatives of the saccule vary greatly in the different vertebrate classes. In teleosts (bony fishes), amphibians, reptiles, and birds there is a lagena (a curved, flask-shaped structure), with its macula, the macula lagenae. Only the amphibians have a papilla amphibiorum, which is located near the junction of the utricle and the saccule. In some amphibians and in all reptiles, birds, and mammals, there is a papilla basilaris, which is usually called a cochlea in the higher forms, in which it is highly detailed. The elaborate sensory structure of higher types of ears, containing hair cells and supporting elements, is called the organ of Corti.

The macular endings consist of plates of ciliated cells (cells with short, hairlike projections) along with accessory cells, all surmounted by an otolith (a calcareous mass containing numerous particles of calcium carbonate embedded in a gelatinous matrix) or, in teleosts, by one large mass of calcium carbonate. The crista endings contain moundlike groups of sensory cells with supporting cells; the sensory cells have elongated cilia that are embedded in a gelatinous body, the cupula, which forms a sort of valve across an expanded portion of each semicircular canal. The papillae contain plates or ribbons of ciliated cells in a structural framework that lies on a movable membrane, except in amphibians, in which the papillae are on a solid base. These ciliated cells are not surmounted by an otolithic mass or a cupula, but some of the cilia are attached either directly or indirectly to a tectorial membrane (a membrane with one edge fixed to a stationary base, thus anchoring the cilia) or to an inertia body (a mass lying over the ciliated cells and restraining the movements of the cilia).

Functions of labyrinthine endings

The endings have different functions: the macular organs serve primarily as gravity receptors and detectors of sudden movements; the crista organs serve for the perception

From *Biological Review* (1936)

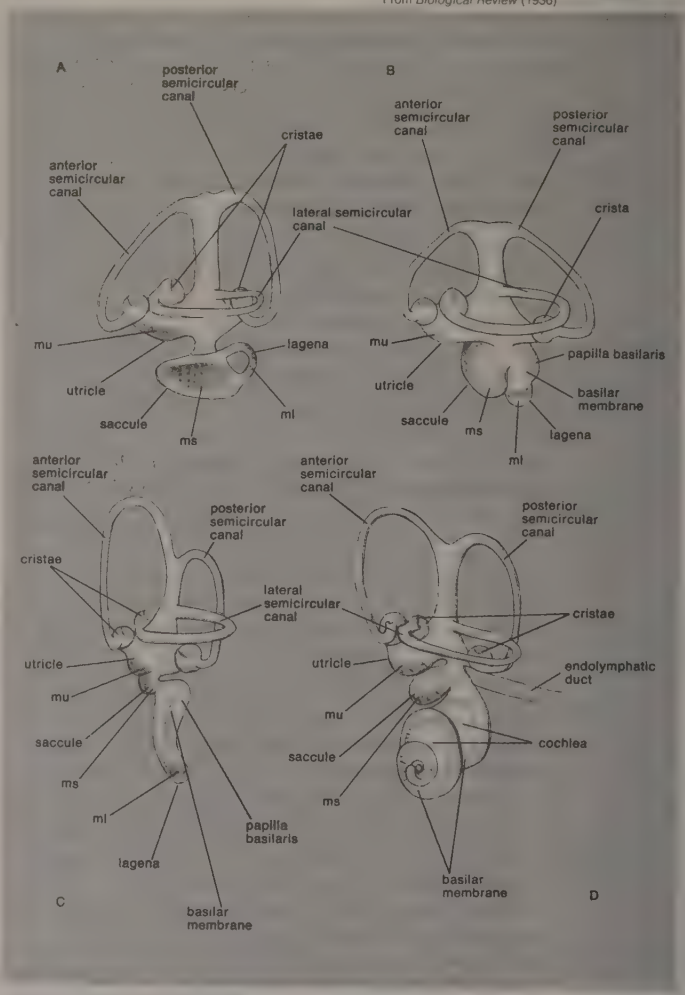


Figure 27: Generalized labyrinth of (A) fish, (B) turtle, (C) bird, and (D) mammal. (Abbreviations: mu, macula utriculi; ms, macula sacculi; ml, macula lagenae.)

of rotational acceleration; and the papillae serve for hearing. As structural relations suggest, the auditory endings are derived either from the other labyrinthine receptors or from the primitive labyrinthine epithelium.

Hearing in fishes. The cyclostomes and the elasmobranchs (*e.g.*, sharks and rays) possess a labyrinth with maculae and cristae but have no auditory papillae. There are, nevertheless, two possible ways by which some of these cartilaginous fishes, especially the sharks, react to sounds in the water: by means of the macular organs and by means of the lateral-line apparatus. (For details on these mechanisms, see above *Mechanoreception: reception of external mechanical stimuli.*) It is in the bony fishes (teleosts) that a true ear whose function is hearing first appears among the vertebrates. This ear, which occurs in a number of forms, has varying degrees of effectiveness as a sound receiver; some fishes hear well, others poorly. The differences arise, at least in part, from the accessory mechanisms that aid in the utilization of sound energy.

The basic auditory mechanisms in teleosts. In most fishes, especially in many marine forms, the auditory mechanism is relatively simple, consisting of macular endings that evidently have been diverted from their primitive functions as detectors of gravity and motion. The important change is not in the structure of the end organ but in its innervation—the nerve supply has connections that transmit auditory information. It is thought that in most teleosts the change to an auditory function has occurred in the saccular macula, and probably the lagenar macula as well, and that the utricular macula continues as a receptor for gravity and motion.

Inertia principle in the teleost ear

The simple macular ending of the teleost ear is stimulated by sound through the operation of an inertia principle. Sound waves pass readily through the water and into the body of the fish, causing most of the tissues to vibrate in a uniform manner. The macular otolith, however, represents a discontinuity; because its density is greater than that of the other tissues, it exhibits an inertia effect (resistance to movement). Its motions not only lag behind those of the surrounding tissues but are probably of lesser amplitude as well. Accordingly, a sound creates a relative motion between the otoliths and the other tissues. More specifically, there is relative motion between the bodies of the hair cells, which rest on a tissue base, and the cilia of these cells, the tips of which are in contact with the otolith. This method of stimulating the auditory hair cells is inefficient, however, because of the relatively small difference in density between the body tissues and the otoliths.

Special stimulation mechanisms. In certain groups of teleosts the efficiency of hair-cell stimulation has been increased by a discontinuity that is nearly 1,000 times greater than the one between tissue and otolith; this is the discontinuity between the otolith and a gas bubble. Although there are varying anatomical methods of achieving it, the simplest arrangement, which is found in clupeids, mormyrids, labyrinthine fishes, and a few others, consists of a gas-filled sac that lies against one wall of the labyrinth. In clupeids (*e.g.*, herring), a group in which the utricular macula rather than the saccular or lagenar maculae has an auditory function, long anterior extensions of the swim bladder form air sacs, one adjacent to each utricular macula. In the mormyrids, which include the elephant-nosed fish, a similar condition exists in early life; during adult development, however, the connections with the swim bladder disappear, leaving the air sacs connected with the saccular and lagenar endings. The gas content of these sacs is then maintained by special glands that extract gas from the blood. Air sacs arise in various other ways.

Air sacs

One large group of fishes, referred to as the Ostariophysi (*e.g.*, catfishes, minnows, and carps), has no air sac adjacent to the labyrinth, but a possibly equivalent condition is achieved through a mechanical connection between the swim bladder and fluid chambers adjacent to the labyrinth. A chain of three or four small bones, known as the Weberian ossicles, extends from the anterior wall of a part of the swim bladder to a fluid-filled chamber called the atrium, which in turn connects by fluid passages with the two labyrinths in the region of the saccule-lagena complex (Figure 28). In this arrangement the discontinuity is between the air of the swim bladder and the chain of ossicles in contact with it; the relative motion arising from sound stimulation is communicated through the ossicular (bony) chain and the fluid channels to the macular endings.

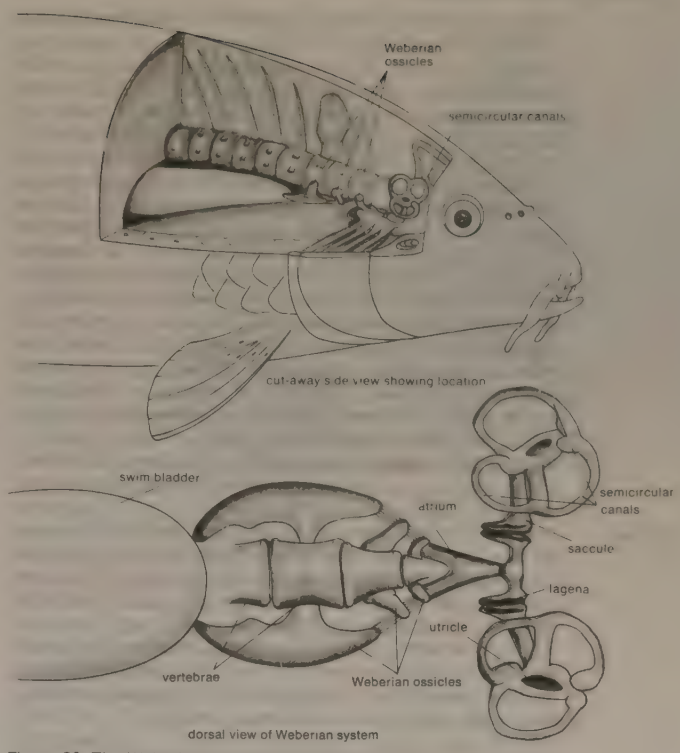


Figure 28: The Weberian ossicles and labyrinths of an ostariophysan fish.
Adapted from *Copeia* (1943) and (bottom) *Nature* (1938)

Regardless of the mechanism employed, however, the ear of all teleost fishes is basically a macular organ. Because it is stimulated by sound that is transmitted to tissues adjacent to the sensory cells and that acts differentially on these cells, this ear is of the velocity type.

Auditory sensitivity of fishes. Although only limited experimental data are available, it appears certain that, in general, fishes with the accessory mechanisms described above have greater sensitivity and a higher frequency range than do those lacking such mechanisms; while upper frequency limits are about 1,000 hertz for many fishes, they are about 3,000 hertz for the Ostariophysi and other specialized types.

Auditory response in goldfish

Many experiments have dealt with the problem of auditory sensitivity in fishes, but the species most extensively tested has been the goldfish, a variety of carp belonging to the Ostariophysi. In one well-controlled investigation, the sound intensities required to inhibit respiratory movements, after conditioning with electric shock, were studied. The greatest sensitivity was found to be around 350 hertz; above 1,000 hertz sensitivity declined rapidly.

In view of the simple anatomical character of the ear, the question of whether fishes can distinguish between tones of different frequencies is of special interest. Two studies dealing with this problem have shown that the frequency change just detectable is about four cycles for a tone of 50 hertz and increases regularly, slowly at first, then more rapidly as the frequency is raised.

Hearing in amphibians. There are three orders of living amphibians: the Apoda, which are legless, wormlike types such as caecilians; the Urodela, which are tailed forms such as mudpuppies, newts, and salamanders; and the Anura, which are tailless forms including frogs and toads. Although members of all three orders have ears, the structures vary greatly in the different groups, and little is known about them except in such advanced types as frogs.

The auditory mechanism in frogs. Although the frog has no external ear (structures on the outside that direct sound vibrations inward), the middle-ear mechanism is well developed. On each side of the head, flush with the

surface, a disk of cartilage covered with skin serves as an eardrum. From the inner surface of this disk, a rod of cartilage and bone, called the columella, extends through an air-filled cavity to the inner ear. The columella ends in an expansion, the stapes, which makes contact with the fluids of the inner-ear (otic) capsule through an opening, the oval window. A second opening in the otic capsule, the round window, is covered by a thin, flexible membrane; it is bounded externally by a fluid-filled space that can expand into the air-filled cavity of the middle ear. When the alternating pressures of sound waves cause the eardrum to vibrate, the vibrations are transmitted along the columella and through the oval window to the inner ear, where they are relayed to the round window in a path across the otic capsule by movements of the inner-ear fluids. Along this path are two auditory endings, the amphibian and basilar papillae, the sensory hair cells of which are stimulated by the fluid movements. These movements are transmitted to the ciliary tufts of the sensory cells by a tectorial membrane, which is suspended from the hair cells in such a way that it can be moved by the oscillations of the inner-ear fluids.

Papillae

As sense organs for hearing, the papillae, which appear for the first time in amphibians, have cells like those in lower vertebrates that serve the same purpose. There are two types of papillae: the amphibian papilla, which is found in all amphibians, and the basilar papilla, which is found in some amphibians. Because they are located in different places in the inner ear, the papillae probably represent two distinct evolutionary developments. Moreover, they operate on a mechanical principle found in no other animal group: a tectorial membrane, moving in response to sound vibrations that have been transmitted to it by the inner-ear fluids, stimulates the sensory hair cells directly through connections to the cilia of these cells. In all higher types of ears, on the other hand, the sensory cells themselves are set in motion by the sound vibrations, while the tips of the ciliary tufts are restrained in one of several ways.

Auditory sensitivity of amphibians. Although it is presumed that all amphibians possess hearing of some kind, the evidence is sparse; only salamanders other than anurans have been studied experimentally. Salamanders trained to come for food at the sound of a tone responded only at low frequencies, up to 244 hertz in one specimen and to 218 hertz in three others.

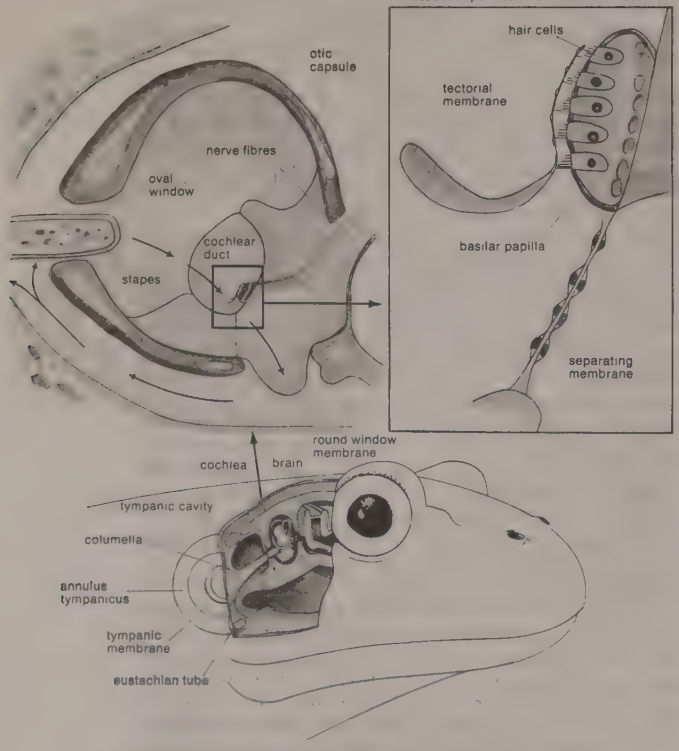


Figure 29: The inner ear of a frog, showing one of the papillae. The arrows indicate the path along which sound vibrations are transmitted by movements of the inner-ear fluids. At the right is an enlarged view of the separating membrane and the basilar papilla.

Frogs, which are of special interest because they first live in the water as tadpoles and then undergo a metamorphosis that equips them for life on land, have been studied more extensively. Considerable modifications of the middle-ear mechanism occur during metamorphosis. Presumably, the tadpole larva has an aquatic ear that is later transformed into an aerial type.

Croaking

Interest in the hearing of adult frogs has been stimulated by their active and often loud croaking during the breeding season. Evidently, their vocalizations assist in the location and selection of mates. The first experimental study of auditory sensitivity in frogs, carried out in 1905, showed that leg movements in response to strong tactual stimuli may be enhanced or even inhibited by sounds.

Somewhat later, following some unsuccessful attempts to train frogs to make behavioral responses to acoustic stimuli, two other methods were employed to determine the sensitivity and range of their hearing. One of these was the recording of changes in the electrical potentials of the inner ear and auditory nerve; the other was the observation of changes in the potentials of the skin (electrodermal responses) to acoustic stimuli. As a result of these investigations, inner-ear potentials and electrodermal responses in the bullfrog have been recorded over a range from 100 to 3,500 hertz. In the treefrog, these same responses have been found in a range that extended from 50 to 3,000 hertz, with the greatest sensitivity from 600 to 800 hertz, and again at 2,000 hertz.

The recording of impulses from single fibres in the auditory nerve of bullfrogs and the green frog indicates that two types of auditory nerve fibres are present. This has led to the suggestion that they represent the different characteristics of the amphibian and basilar papillae. It is believed that the amphibian papilla is more sensitive to low tones and that the basilar papilla is more sensitive to high tones.

AUDITORY STRUCTURES OF REPTILES

The living reptiles belong to four orders: the Squamata (lizards, snakes, and amphisbaenians), the Rhynchocephalia (one rare species, the tuatara of New Zealand), the Chelonia (turtles), and the Crocodylia (crocodiles and alligators). The reptile ear has many different forms, especially within the suborder Sauria (lizards), and variations occur in all elements of its structure—the external ear is often absent or may consist of an auditory meatus (passage) of varying length; the middle ear shows several forms in the different groups; and the inner ear varies in the degree of development of the auditory papilla and also in the ways by which the sensory cells are stimulated by sound.

Lizards. *Auditory structure.* There are about 20 families of lizards, ranging from the chameleon, a divergent type, to the gecko, certain species of which have the most highly developed ears found in the group. The chameleons, of those species studied thus far, have only a few sensory hair cells (40 to 50) in the auditory papilla. The geckos, on the other hand, have several hundred hair cells, and the *Gekko gecko* has about 1,600, the largest known number of hair cells in any saurian. Other lizard species fall between these two extremes in inner-ear development, with the iguanids, the most common lizards in the Western Hemisphere, having from 60 to 200 hair cells, according to the species.

Middle ear of lizards

What may be regarded as the standard type of middle-ear structure in the lizards consists of a tympanic membrane and a two-element ossicular chain that extends from the inner surface of this membrane to the oval window of the otic capsule. As shown in Figure 30, the ossicular chain is made up of two parts: the osseous (bony) columella, whose expanded innermost end (the stapes) fills the oval window, and the extracolumella, a cartilaginous extension that usually spreads out in two to four processes that are embedded in the fibrous layer of the tympanic membrane. Geckos have a single middle-ear muscle attached to the lateral part of the extracolumella; evidently, contractions

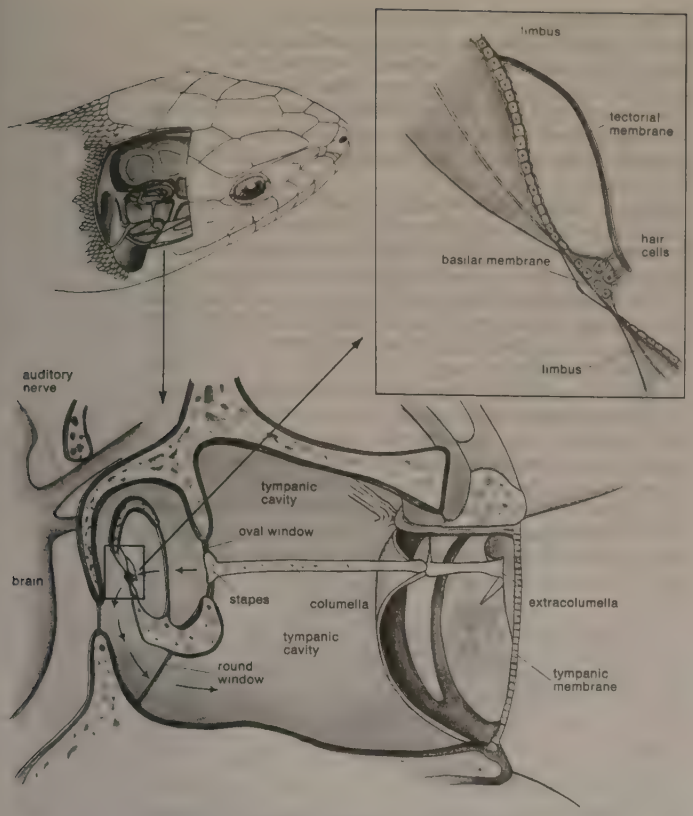


Figure 30: The auditory structures in the right lizard ear as seen from behind and from the left (see text).
From *Journal of Auditory Research* (1965)

of this muscle stiffen the extracolumella, thereby dampening the ossicular motions and protecting the ear against excessively intense sounds.

The auditory part (cochlea) of the inner ear (Figure 30) consists of a basilar membrane lying in an opening in the limbus, which is a plate of connective tissue. The form of the basilar membrane, which is unlike the structure of the same name in amphibians and is clearly of different origin, varies from a simple oval in iguanids to a long, tapered ribbon in gekkonids. In many species the middle portion of the basilar membrane is greatly thickened, especially in some regions of the cochlea. Over this thickening, which is called the fundus, lies the auditory papilla proper—*i.e.*, that part of the cochlea in which the sensory hair cells are held in a framework of supporting tissues and cells. The hair cells usually occur in regular transverse rows, with the number of cells in a row varying along the cochlea. They have a tuft of cilia, the so-called sensory hairs, of graduated lengths, the longest of which are usually attached either directly or indirectly to a tectorial membrane. This membrane arises from a region of the limbus that is usually elevated, often strikingly so, and runs as a thin web or sheet to the region of the hair cells. Only rarely does the free edge of the tectorial membrane connect directly with the cilia of the hair cells; usually there are intermediate connecting structures that take a variety of forms, from simple fibres to relatively massive plates.

Modes of auditory stimulation

The function of the tectorial membrane and its connections to the ciliary tuft of a hair cell is to immobilize the tuft when the body of the hair cell moves in unison with the basilar membrane on which it rests. This produces a relative motion between the ciliary tuft and the body of the cell and stimulates the cell. All auditory stimulation depends ultimately upon this relative motion, and the means just described for achieving it can be regarded as the most fundmental process by which sounds are perceived. Although it is employed in the great majority of ears, it is not the only mode of stimulation. Another mode is that in the ears of fishes, in which an otolith lies upon the ciliary tufts and, by its inertia, reduces and alters the motion of the tuft relative to the cell body. Still another method is the one in the frog papilla, in which the tectorial membrane is moved by the cochlear fluids while the body of the sensory cell remains at rest.

In some lizards the inertia principle has a form different from that found in fishes. In the former, a body called a sallet lies upon the ciliary tufts of a group of hair cells and, by its inertia (or by an equivalent means), restrains the movement of the cilia when the cell body is made to move. The result is a relative motion and a stimulation of the hair cells, like the more common restraint by a tectorial membrane.

The ears of two lizard families show only the inertial restraint method of stimulation; in several other families this method functions in some regions of the cochlea for certain hair cells. Hair-cell stimulation by two or more different arrangements within the same cochlea, however, is the rule rather than the exception because of its many advantages. Although the tectorial-restraint method provides great sensitivity for individual cells, the sallet system also attains good sensitivity, but in another way: by causing many cells—those in common contact with a given sallet—to work in parallel, thus producing a spatial summation. The sallet system has the advantage of being more resistant to damage by overstimulation from intense sounds. In such lizards as the geckos, for example, in which the hair cells are divided nearly equally between tectorial and sallet systems, an exposure to excessive sound has been observed to break all the tectorial connections to the hair cells while leaving the sallet connections intact. But even though the most sensitive hair cells are inoperative, the animal can respond to sounds, although with lesser acuity.

Hearing abilities of lizards. The lizards are the lowest vertebrates to have a well-developed spatial differentiation of the cochlea in which different regions respond to different frequencies of tone. The problem of tonal discrimination has been somewhat solved in frogs, in which the differential responses to tones by the two papillae may provide some information concerning the pitch of sounds. The mechanism in frogs, however, is a poor one, as it can give only crude and uncertain cues at best.

Tonal discrimination

In some lizards, such as iguanids and agamids, a minimum of structural variation occurs along the cochlea; in others (*e.g.*, geckos, which have very extensive differentiation along their extended basilar membranes) the differentiation is almost as great as that in higher vertebrates, including man. Most geckos are nocturnal in habit and use vocalizations to maintain individual territories and probably to find mates.

Although it has been possible to train two species of lizards (*Lacerta agilis* and *Lacerta vivipara*) to make feeding movements in response to a variety of sounds, including tones between 69 and 8,200 hertz, most attempts to train lizards to respond reliably to tonal stimuli have failed. The one useful method thus far developed to study the sensitivity of these animals to sounds involves recording electrical responses in the ear and in the auditory nervous system. Although such observations have provided information about peripheral response to sounds, they do not reveal anything about other processes in the nervous and behavioral systems.

Electrical responses in the cochlea of many lizard ears show considerable variations: in absolute sensitivity, in the tonal regions in which responsiveness is best, and in the extent of the frequency range. It has been concluded that most lizards have good auditory sensitivity over a range from 100 to 4,000 hertz and relatively poor hearing for lower and higher tones. This auditory range is not very different from that of man, although somewhat more restricted than that of most mammals.

Snakes. Without much doubt, snakes developed from some types of early lizards but lost their legs when they adopted habits of burrowing in the ground. Although some snakes burrow, others have taken up different habits: many species live on the surface of the ground, several are largely aquatic, and some live in trees. All, however, show drastic ear modifications that reflect their early history as burrowers; for example, there is no external ear—*i.e.*, no opening at the surface of the head for the entrance of

Auditory modifications of snakes

sound. This fact, together with a seeming indifference to airborne sounds, has led to the supposition that snakes are deaf or that they can perceive only such vibrations as reach them through the ground on which they crawl.

This supposition is incorrect; snakes are sensitive to some airborne sound waves and are able to receive them through a mechanism that serves as a substitute for the tympanic membrane. This mechanism consists of a thin plate of bone (the quadrate bone) that was once a part of the skull but that has become largely detached and is held loosely in place by ligaments. It lies beneath the surface of the face, covered by skin and muscle, and acts as a receiving surface for sound pressures. The columella, attached to the inner surface of the quadrate bone, conducts the received vibrations to its expanded inner end, which lies in the oval window of the cochlea. If the columella is severed, the sensitivity of the ear is significantly reduced.

Although the sensitivity of the snake ear varies with the species, it is appreciably sensitive only to tones in the low-frequency range, usually those in the region of 100 to 700 hertz. For this low range the large mass of the conducting mechanism and the presence of tissues lying over the quadrate bone are not of any great consequence. Moreover, while the sensitivity of most snakes to the middle of the low-tone range is below that of most other types of ears, it is not seriously so. In a few snakes, however, the sensitivity is about as keen as in the majority of lizards with conventional types of ear openings and middle-ear mechanisms.

That the ears of the snake receive some aerial sound waves instead of depending exclusively on vibrations transferred from the ground has been proved by recording the potentials in the cochlea of one ear while rotating the animal in front of a sound-wave source so that the ear being studied was sometimes facing the source and sometimes directed away from it. The recorded potentials were significantly greater when the ear was facing the source. There would have been no difference in the responses if the sound first set up vibrations in the ground and these were then transmitted to the body. This observation also shows that the ears of the snake can determine the direction of a sound in terms of its relative intensity in the two ears. Although snakes can perceive vibrations from the ground that are present at a sufficient intensity, this ability is not peculiar to them; all ears respond to vibrations transmitted to the head.

Amphisbaenians. The amphisbaenians form a little-known group of reptiles. Because they are burrowers and live almost entirely underground, they are seldom seen. The one species in the United States, *Rhineura floridana,* is found in some parts of Florida; a number of species occur in other regions of the world, especially in South America and Africa.

The animals construct a maze of underground tunnels, which they patrol in search of such food as grubs and worms. Although small eyes below the body surface can receive light through a transparent scale, amphisbaenians evidently make little use of vision. There is reason to believe, however, that they use hearing to locate their prey.

Amphisbaenians, like snakes, have no surface indication of an ear; a receptive mechanism below the surface and different from that in snakes conveys vibrations to the inner ear. In the oval window, which occupies the entire lateral surface of the otic capsule, is a stapes. The head of the stapes in most species is directed laterally and forward; it is united by a joint with a rod of cartilage (the extracolumella) that extends forward along the face, in the line of the lower jaw. The extracolumella lies below the surface, where it makes close contact with and finally enters a dense layer of the skin. When the facial region is exposed to sounds, the vibrations are transmitted through the dense layer of the skin to the extracolumellar rod and then through it to the stapes, finally reaching the fluid of the inner ear. That this is the route of sound conduction has been proved by cutting the extracolumella at different places and observing the reduction of recorded responses in the ear.

Auditory structures in the face of amphisbaenians

The auditory mechanism of amphisbaenians varies somewhat according to species but is substantially as described above. The sensitivity, which also varies with species, is surprisingly high in some species, considering the unusual nature of the mechanism involved. Studies similar to those described for snakes have proved that this ear receives aerial sounds and that it can determine the direction from which the sound originated. As expected, this ear also responds to mechanical vibrations communicated directly to the skull.

Turtles. It is sometimes supposed that the turtle's ear is a degenerate organ, largely or even completely unresponsive to sound. Although the turtle's ear is unusual in some respects, and can be regarded as specialized in its manner of receiving and utilizing sounds, it is not a degenerate organ. There is good evidence that turtles are sensitive to low-frequency airborne waves and that some species have excellent acuity in this range.

A plate of cartilage on each side of the head serves as a tympanic membrane. Leading inward from the middle of this plate is a two-element ossicular chain consisting of a peripheral extracolumella and a medial columella the expanded end (the stapes) of which lies in the oval window of the otic capsule. Within the otic capsule are the usual labyrinthine endings, including an auditory papilla. As shown in Figure 31, the auditory papilla lies in a

From *Proceedings of the National Academy of Sciences* (1956)

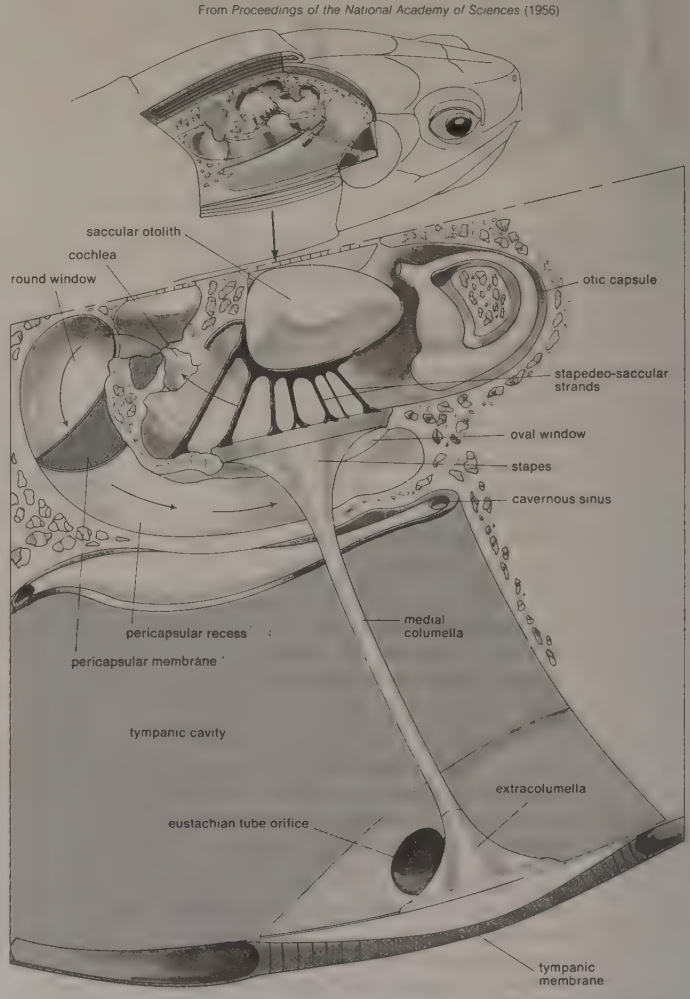


Figure 31: *Structure of the turtle ear.*
(Top) View of the head, showing the location of the otic capsule; (bottom) detail of the otic capsule. Arrows indicate the path along which vibrations are transmitted by movements of the cochlear fluids.

path between the oval window and an opening (the round window) in the posterior wall of the otic capsule. Unlike the round window in most ears, that in turtles has no membranous covering for transmitting pressure changes to the air-filled cavity of the middle ear. Instead, the opening leads to a fluid-filled chamber, the pericapsular recess, that extends laterally and anteriorly to enclose the external

portion of the stapedial expansion of the columella. A pericapsular membrane separates the perilymph (fluid) of the otic capsule from the fluid of the recess. When the stapes is moved inward by the columella at one phase of a sound vibration, the fluid of the otic capsule is displaced, causing a pressure change that, after passing through the sac containing the auditory endings, continues in a circuitous course to the external surface of the stapes. This circuit is indicated by the arrows in Figure 31. When the columella moves outward, the fluid circuit reverses itself. Hence the result of a continuous sound wave is a surging back and forth of the fluids in the otic capsule and the pericapsular recess at the same frequency as that of the sound.

The special mechanical arrangement in the turtle ear is fully effective within the low-frequency range. Indeed, the relatively large mass of tissue and fluid involved in the response to sounds is in part responsible for the efficiency of the ear at low frequencies and also for the rapid loss of sensitivity as frequency increases.

This type of cochlear response to sounds is not peculiar to turtles; it is also found in snakes, through a structural arrangement of similar form. Although it also occurs in amphisbaenids, the fluid path in these animals is entirely different: it proceeds through the perilymphatic recess into the brain cavity and then by an anterior passage across the head to the lateral surface of the stapes.

Certain experiments involving the turtle's sensitivity to sounds have used training methods (conditioned responses); only a few have met with success. It has been found that turtles of the species *Pseudemys scripta*, trained to withdraw their head, respond to sound over the low-frequency range, with the greatest sensitivity in the region of 200 to 640 hertz. This result is in close agreement with electrophysiological observations in which it has been found that impulses could be obtained from the auditory nerve of *Chrysemys picta* for tones between 100 and 1,200 hertz, with highest sensitivity for tones below 500 hertz. Similar results have been obtained by additional observations of this kind with several other species of turtles, some of which are very sensitive to a narrow band of frequencies in the low-tone range. Evidently, the type of receptor mechanism in the turtle can achieve great sensitivity through mechanical resonance at a particular region of the low-frequency scale.

Low-frequency sensitivity in turtles

Evidence has also been obtained that these responses are to aerial waves and not to vibrations set up in the ground. The sensitivity to surface vibrations was considerably poorer than that to aerial sounds. In addition, cutting the columella seriously impaired the responses to aerial sounds but hardly affected responses to mechanical vibrations applied to the turtle's shell.

Crocodiles. The order Crocodylia includes four groups of closely related forms: crocodiles, alligators, caimans, and gavials. The crocodile ear, although clearly reptilian in general structure, has a number of peculiar features. Leading to a tympanic membrane on each side of the head is a shallow external passage the outside opening of which is protected by an earlid that is closed when the animal enters the water and dives. Beyond the tympanic membrane is a middle-ear cavity, with the one on the right connected to the one on the left by an air passage that runs across the head above the brain. A sound presented to one ear, therefore, reaches the other ear about equally well. A columellar system connects the tympanic membrane to the oval window of the otic capsule, as in other reptiles. The inner ear is highly developed and bears many similarities to the cochlea of birds, described in the next section. Elongated and slightly curved, the cochlea contains about 11,000 sensory hair cells, about seven times as many as found in that of the most advanced lizard (*Gekko gecko*).

In comparison to some lizards, the cochlea of *Caiman crocodilus*, which has been most extensively studied, exhibits only a moderate degree of structural differentiation. Yet in this cochlea fibre bundles that extend from the root portion of the tectorial membrane separate into fine fibres that form individual connections with the ciliary tuft of each hair cell. This arrangement is not a common one, though present in certain lizards, such as the chameleons, and also in some degree in birds. It probably provides a high level of specificity in the stimulation process or as much specificity as the overall mechanical pattern permits.

The hearing of crocodilians has not been studied very extensively. It has been noted that the breathing rate in a crocodile accelerates in response to loud sounds, such as the firing of a gun, and it has been observed that specimens of the Mississippi River alligator produce vocalizations of roaring or hissing when low-frequency sounds are made by blowing a horn or by plucking a metal rod. Studies of the electrical potentials in the ear of *Caiman crocodilus* show that it is sensitive to frequencies ranging from 20 to 15,000 hertz.

HEARING IN BIRDS

The avian auditory structure. Ears of birds show considerable uniformity in general structure and are similar in many respects to those of reptiles. The outer ear consists of a short external passage, or meatus, ordinarily hidden under the feathers at the side of the head. Most birds have a muscle in the skin around the meatus that can partially or completely close the opening.

The tympanic membrane bulges outward as in most lizards. In the songbirds, however, it consists of two separate membranes, with the outer one apparently serving to protect the inner one from injury. From the inner surface of the tympanic membrane an ossicular chain transmits vibrations of the cochlea. As in lizards, the chain consists of an osseous inner element, the columella, and a cartilaginous extracolumella that extends the columella peripherally and connects with the tympanic membrane.

The tympanic membrane of birds

The cochlea of birds is similar to that of crocodiles, consisting of a short, slightly curved bony tube within which lies the basilar membrane with its sensory structures. The length of the basilar membrane varies between 2.5 and 4.5 millimetres (0.1 and 0.2 inch) in most birds, but in the owls it may reach 10 millimetres (0.4 inch) or more. At the end of the cochlea is another ending with a different function, the lagena and its macula.

Auditory sensitivity in birds. Using the conditioned-response method to study auditory sensitivity in a small songbird, the bullfinch, responses over a frequency range from 100 to 12,800 hertz have been observed. The electrophysiological method was first applied to the study of hearing in birds in 1936. In this study impulses from the cochlea of pigeons were recorded for tones usually up to 10,000 hertz and occasionally as high as 11,500 hertz. Although this method has been used since 1936, few detailed and quantitative results have been obtained; nevertheless, one striking characteristic revealed by these studies has been the high degree of sensitivity in the low and middle range and the very rapid decrease in the high tones.

Uses of hearing in birds. Like other animals, birds use hearing to warn them of enemies and other kinds of danger. To a degree hardly equalled in lower species, they also use hearing in social relations and communication. Many male birds sing to hold their territories and to attract mates. Some birds also use vocalizations to identify their mates or group members. During the breeding period of the emperor penguin, for example, the male leaves his mate for a journey taking many days in order to obtain food. Upon returning to the general area where his mate has remained with a pack of hundreds of birds, the male is able to locate and to recognize his partner by an interchange of calls.

There is good reason to believe that certain birds, including the swiftlets (*Collocalia*) of Asia and Australia, the oilbirds (*Steatornis*) of tropical America, and possibly a few others, are able to use echolocation when flying in the dark caves that they inhabit. Moreover, it is well established that many owls locate and catch their prey by auditory cues. On a dark night, an owl perched in a tree can hear the rustling sounds made by a mouse in the grass and leaves on the ground below; by accurately localizing this signal, he can make his strike and capture the prey without any visual aid.

HEARING IN MAMMALS

Auditory structure of mammals. In the mammals the ear reaches its highest level of development, with well-

differentiated divisions of outer ear, middle ear, and inner ear. Except in some of the sea mammals, in which certain modifications and degenerations have taken place, these structures carry out their functions in a remarkably regular manner.

The pinna

The outer ear consists of pinna (or auricle) located behind the ear opening and partially enclosing it and an auditory meatus that leads inward. The pinna varies greatly in size relative to the size of the animal, being large enough in many species to serve a useful purpose in the collection and reflection of sounds. Many mammals can move the pinna back and forth to regulate in some degree the entrance of sounds to the auditory meatus, which transmits the sounds inward to the tympanic membranes. In some mammals, such as many of the marine types, the external opening can be closed to keep out water when the animal dives, and in certain species of bats the tube itself contains a valve that can be closed to protect the ear against undesirable sounds.

The middle ear of mammals consists of a tympanic membrane, an ossicular chain of three elements, and two tympanic muscles. The tympanic membrane bulges inward, unlike the usually outward-bulging membrane of reptiles and birds. The elements in the ossicular chain are the malleus (hammer), incus (anvil), and stapes (stirrup), so named because of the resemblance of the bones to these objects. The malleus is attached to and partly embedded in the fibrous layer of the inner surface of the tympanic membrane. It connects to the incus, which connects in turn to the stapes, the footplate of which lies in the oval window of the cochlea.

One tympanic muscle extends from an attachment to the skull to an insertion on the malleus. Another muscle has its insertion on the neck of the stapes. By their contractions, both muscles add friction and stiffness to the ossicular chain, thereby reducing its mobility and protecting the inner ear from excessive sounds. The contraction of the muscles is a reflex action and occurs in both ears at the same time in response to loud sounds.

The coiled cochlea

The inner ear is called the cochlea because in man this structure is a complex tube coiled into about 2.5 turns, thus bearing some resemblance to a snail's shell, from which the term is derived. The name cochlea has now been extended to include the auditory portion of the labyrinth in all animals, even when the structure is not coiled, as in reptiles, birds, and egg-laying mammals. In the mammals in which it is coiled, the number of turns in the cochlea varies with species from a little less than two to as many as four. The guinea pig and its relatives have the largest number of cochlear turns. Extending along the inside of this coiled passage is the basilar membrane, bearing on its surface the sensory structure known as the organ of Corti, which contains the hair cells.

In mammals a uniform system is employed in the stimulation of the hair cells by sounds. A relatively thick tectorial membrane, anchored securely on one edge to the supporting structure (the limbus), lies with its free portion over the hair cells and with the cilia of these cells firmly attached to the lower surface of this portion. When vibratory movements of the basilar membrane cause the bodies of the hair cells to move, the tips of the cilia are restrained by their attachments to the tectorial membrane. Hence the relative motion between the bodies and cilia of the hair cells stimulates them.

The sizes, shapes, and spatial relations of many otic structures vary in the different mammalian species, but it is thought that the same basic principles of operation are involved. This uniformity contrasts with their situation in reptiles, in which different systems are present both in different species and sometimes within one ear.

Auditory sensitivity among mammals

A number of features are of particular significance in determining the sensitivity and frequency range, which vary with species. Because large masses involve great resistances when moved at high frequencies, the size and mass of the moving parts determine to some degree the variations of sensitivity with frequency and the frequency limits within which the ear operates. The ossicular chain is a mechanical lever, and its lever ratio and the difference in area between the tympanic membrane and the stapedial footplate determine the efficiency of sound transmission from air to the cochlear fluid. The mechanical characteristics of the cochlea and the degree of variation of these characteristics along its extent determine the frequency range of hearing and the degree to which different tones can produce different response patterns. Finally, the numbers and distribution of hair cells along the basilar membrane and the density and specificity of innervation of these cells determine the delicacy and precision with which their periodic activity and spatial patterns are registered by the central areas of the auditory nervous system.

These anatomical features have been studied in detail in a few animals: among the mammals, mainly in cats, guinea pigs, and to a lesser degree in man. The functional aspects, as shown in responses to sounds and to discriminations among different sounds, have been considered principally in man and to a much more limited extent in other mammals. (The characteristics of human hearing are treated at length below; see *Human hearing.*) Some of the auditory characteristics of mammals below man are described in the sections that follow.

Hearing in subhuman mammals. *Primates.* The hearing of other species in the division of mammals to which man belongs has always been of special interest. A number of species have been studied, including monkeys, marmosets, and chimpanzees among the primates considered as the most advanced, the anthropoids; and tree shrews, lemurs, and lorises among the more primitive.

By using a variety of training methods with chimpanzees, monkeys, and marmosets, behavioral thresholds have been recorded in response to sounds of different intensities and frequencies. When compared with each other and with man, it has been found that the hearing sensitivity of these animals and man is remarkably similar over a range of frequencies from 100 to 5,000 hertz, after which the sensitivity begins to differ. The differences observed at the higher frequencies, however, may be partly attributed to variations in experimental procedures. Thus, the results for the chimpanzee stop at 8,192 hertz because this was the highest tone used in the tests. Other observations have shown that chimpanzees can hear tones up to about 33,000 hertz and that young human subjects often hear tones as high as 24,000 hertz. It is also evident that monkeys and marmosets of the species studied can hear still higher tones.

Common laboratory animals. Certain mammals have long been favourite subjects for various kinds of biological studies in the laboratory, largely because of their convenient size, hardiness under caged conditions, and gentle temperament. Familiar among these are cats, dogs, guinea pigs, rats, mice, rabbits, and, more recently, hamsters, chinchillas, and gerbils. Auditory sensitivity functions have been obtained in these animals by a variety of behavioral and electrophysiological methods.

When measured behaviorally by conditioned responses and then plotted on a curve, the auditory threshold sensitivity of cats, guinea pigs, and chinchillas is much the same—a progressive improvement in sensitivity as the frequency is raised until the middle tones (about 500 to 5,000 hertz) are reached, at which point sensitivity tends to remain the same, and then shows a rapid loss in the upper frequencies. There are differences, however, in the maximum sensitivity attained in the middle region, with the guinea pig the least sensitive and the cat the most sensitive of the three species.

Differences between behavioral and electrophysiological observations

Sensory responses in the cochlea of mammals have been measured electrophysiologically by placing an electrode on the round window membrane. Unlike behavioral curves, however, the curves obtained by plotting the sound required to produce an arbitrary amount of electrical potential of the cochlea do not represent auditory thresholds. Instead, their usefulness is largely in their shapes, which indicate in a relative way the regions of good and poor sensitivity. In addition, these curves represent the performance of the peripheral portion of the auditory mechanism up to the point at which the sound stimulus activates the sensory hair cells in which the potentials are generated. Hence, unlike the curves obtained by behavioral responses, those obtained by cochlear potential methods do not indicate

the performance of the central auditory nervous system (the nerve connections between the ear and brain and those parts of the brain in which neural impulses from the ear are processed to produce behavioral responses).

In the simpler animals, the two types of curves are much alike, judging from the very limited evidence available. In mammals, however, the behavioral curves differ from the cochlear potential curves in three ways. In the behavioral curves there is (1) an exaggerated gain in sensitivity to tones of low frequency, (2) a greater sensitivity to the medium-high tones, and (3) a more rapid loss of sensitivity to the extreme-high tones and a lower frequency of the upper limit. These differences are believed to arise mainly through the elaborate neural processing that takes place in the more highly developed mammalian nervous system, a processing that improves the sensitivity to high-frequency tones but reaches a limit of effectiveness and finally fails above some frequency limit. With these conditions in mind, the electrophysiological curves can be used to predict reasonably well an animal's behavioral responses to sound waves.

Large mammals. Because most of the mammals in which hearing has been studied by laboratory methods are small, much less is known about the auditory capabilities of large ones, even of such domesticated animals as horses and cows; nevertheless, it is usually assumed that the auditory capabilities of these animals are much like those of humans. At least they hear sounds in man's vocal range because they seem to respond to verbal signals. Elephants, for example, trained as working animals, are said to obey as many as 30 different commands. A number of wild animals of medium and large size—raccoons, opossums, and several members of the cat and dog families—have been studied electrophysiologically by the cochlear-response method. Their sensitivity curves are fairly similar in form and in the upper limits attained.

Marine mammals. Of special interest are the sea mammals, which have been derived from early land species and which have undergone certain changes in order to adapt themselves to at least a partially aquatic existence. In the course of adapting to marine conditions, however, some sea mammals, such as seals and sea lions, seem to have made only limited alterations in their ear structures. In addition to being able to close the meatus when diving, their pinnas have been greatly reduced or essentially lost, a feature of streamlining for rapid progress through the water.

Adaptations for aquatic hearing

There are three possible ways that the hearing of marine mammals might be adapted to an aquatic environment: (1) unchanged aerial hearing, with no aquatic adaptation; (2) conversion to an aquatic type of hearing with loss of good hearing for aerial sounds; and (3) development of some kind of double system, with at least serviceable reception of both aerial and aquatic vibrations. In a study of hearing in the common seal, in which responses to aerial and aquatic stimuli were compared, it was found that this animal has a greater sensitivity to aquatic sounds, especially in the upper frequencies, which extended to the remarkably high frequency of 160,000 hertz. Yet, although the seal has made an adjustment for hearing in water, it has not sacrificed the quality of its aerial hearing, which remains at an excellent level, especially for one frequency around 2,000 hertz and another around 12,000 hertz. These differences in auditory senstivity suggest that the mechanisms in this animal for aerial and aquatic hearing are somehow different, but no complete explanation of the adaptations has yet been found.

Whales, on the other hand, have converted their ears to a truly aquatic form, apparently with some sacrifice of aerial reception. The study of their ears and hearing has been carried out in only a few species of the toothed whales, which produce sounds and use their ears in the process of echolocation (see next section).

The ear of whales has undergone extensive changes. The pinna is absent and the external ear opening has been reduced to such a minute size, almost a pinhole in some species, that it no longer serves as a path for the entrance of sound. The eardrum, although present in a modified form, seems to serve no useful purpose; it is connected to the malleus only by a ligament, and this connection can be cut without an ensuing loss of sound reception. The usual three ossicles of the middle ear are present, with the footplate of the stapes resting in the oval window. These ossicles are much more massive than the ordinary mammalian ossicles.

It appears that the whale ear has been converted to a true aquatic type, functioning according to principles similar to those found in the ears of fishes, as described earlier. Sound vibrations in the water readily pass through the tissues of the head and reach the deep-lying middle- and inner-ear structures. Probably the ossicles represent an inertial mass in somewhat the same way that the otolithic body does in fishes. Because of their inertia, the ossicles tend to move with smaller amplitudes and in different phase relations when the tissues of the head, including parts of the cochlea, are set in vibration. This difference in relative motion produces an alternating displacement of the cochlear fluid, which is in contact with the footplate of the stapes and which can be set in motion because of the presence of a pocket of gas in the region of the round window. The performance of the whale ear has been measured in an exact manner throughout the frequency range in one species, the bottle-nosed dolphin (*Tursiops truncatus*). By a conditioned-response method, it has been found that this animal possesses excellent auditory sensitivity that extends well into the high frequencies.

Echolocation in bats. Bats are divided into the large bats and the small bats. With one or two exceptions, the large bats live on fruits and find their way visually. The small bats feed mostly on insects, catching them on the wing by a process known as echolocation. As was mentioned earlier, echolocation is a process in which an animal produces sounds and listens for the echoes reflected from surfaces and objects in the environment. From the information contained in these echoes, the animal is able to perceive the objects and their spatial relations.

How echolocation works

Bats produce sounds with the larynx, an organ in the throat that has undergone certain adaptations that make it unusually effective in producing intense, high-frequency sounds. The character of the sounds varies with the species and also with the particular activity. On striking a small object such as a flying insect, the emitted sounds are reflected with only a small fraction of their original energy; the sound is further weakened before reaching the ears of the bat when it must travel some distance through the air.

Although the frequency of bat cries varies with species, their cries usually occur in a range between 80,000 and 30,000 hertz. In most species, such as *Myotis lucifugus* and *Eptesicus fuscus*, the cry is a frequency-modulated pulse of sound; it begins at a high frequency, say, of 70,000 hertz, and in about 0.2 second declines in frequency to about 33,000 hertz. The starting frequency may vary, even in successive cries; a second pulse might begin at 60,000 and end at 30,000 hertz. The greatest energy in the cry is usually in the middle of this frequency range, perhaps around 50,000 hertz in the species mentioned above.

The use of such high frequencies is an essential feature of the bat's sonar system. In order to determine the nature of objects by reflected sounds, it is necessary that the wavelengths of the sound be small in relation to the dimensions of the objects—indeed, as small as possible if fine details are to be represented.

An important problem of echolocation is how the bat is able to detect reflected sounds, often in the presence of disturbing noises, and to obtain the information necessary for tracking and catching an insect as well as discriminating between this object and others in the environment. This problem involves considering first the structure of the auditory mechanism in bats and then the nature of their hearing.

The external ear of bats is usually well developed. In most species the pinna is large relative to the size of the head, and in those species called the whispering bats, because they make such faint sounds, this structure is huge. With its large surface, the pinna acts as an efficient collector and resonator of high-frequency sounds. It is also freely movable and can be rotated and inclined in various ways. The meatus leads inward to the eardrum and, as already

mentioned, contains a valve that can be closed to reduce the entrance of sounds. The middle ear of bats is of the usual mammalian pattern—a three-part ossicular chain—but its structure is impressive in the extraordinary delicacy of the moving parts. The two tympanic muscles, however, are relatively large.

The cochlea of bats also shows the general mammalian form, but there are variations that may be significant for the special functions that are performed by this ear. The basilar membrane is not particularly well developed; it is short in comparison with that of most mammals, and its structural variation from basal to apical ends is only moderate in extent. Whereas most basilar membranes are rather strongly tapered in width, being narrow at the basal end of the cochlea and several times broader near the apical end, in the bat there is only a slight taper, between twofold and threefold. Another curious feature in the cochlea of bats is the presence of local thickenings of the basilar membrane that may add to the stiffness of the cochlear structure.

The auditory portion of the nervous system has undergone extraordinary development in bats. The regions concerned with hearing are relatively enormous, which is in accord with the great predominance of hearing over the other senses in this animal.

Auditory sensitivity of bats

The hearing of bats has been studied by both electrophysiological and behavioral methods. In the species *Myotis lucifugus,* electrophysiological measurements of cochlear potentials indicate that response is poor in the low frequencies but improves fairly steadily until the range of 2,000 to 5,000 hertz is reached, at which it tends to level off. Beyond 15,000 hertz there are many irregularities but, in general, the sensitivity declines at a rapid rate. The results of similar studies on a specimen of *Eptesicus fuscus* are much the same as those for *Myotis,* though the observations were not extended into the lowest frequencies. The most sensitive range for this species is around 4,000 to 15,000 hertz, after which there is a fairly rapid decline in the upper frequencies.

The behavioral threshold curve for *Eptesicus* has a markedly different form. There is a rapid improvement in sensitivity from 2,500 to 10,000 hertz, but the greatest sensitivity is in two peak areas, from 10,000 to 30,000 hertz and from 50,000 to 70,000 hertz, with a separation by a moderate reduction around 40,000 hertz.

There are other peculiarities of the behavioral sensitivity of *Eptesicus* to sound stimuli that are of particular interest. The rapid loss of sensitivity to tones around 40,000 hertz may be caused by a failure of neural processing for these tones. The slope of the low-frequency end of the curve is unexpectedly steep, and this appears in a region where the cochlear response is passing through its maximum. Nothing like this has been observed in other animals; it seems to be a peculiarity of the bat.

When the cochlear responses of bats are compared with similar responses in other small mammals—as, for example, the rat—there is a general similarity in the results. The rat, however, has better sensitivity as measured by this method, reaching a level of especially good acuity in the range from 20,000 to 60,000 hertz, the range in which the bat sensitivity falls off rapidly. As mentioned previously, it must be kept in mind that the sensitivity indicated by the cochlear potentials is mediated in the peripheral mechanism, before involvement of the central auditory nervous system. When the behavioral response is considered, however, the contribution made by the bat's central auditory nervous system can be appreciated: the region of greatest auditory sensitivity, extending from 10,000 to 70,000 hertz, is the same region as the frequency of the echolocation cries and the one in which bats have the greatest need of acute hearing.

The failure of these bats to exhibit a behavioral response to tones below a frequency of 10,000 hertz can perhaps be explained also in relation to their peculiar use of hearing. This is a region of frequency that has little or no value for echolocation. More than that, it often contains noises of various kinds that, if heard, might be detrimental to this essential function. It has often been observed that bats are not easily disturbed by extraneous sounds of low frequencies, even extraordinarily intense ones. This peculiarity of hearing in bats may account for their resistance to masking sounds. The slight degree of structural differentiation found in the cochlea of bats may represent another aspect of the limitation of their hearing to that part of the sound spectrum most useful in echolocation. Therefore, it appears that the ear of the bat, which is a rather ordinary type of mammalian structure so far as level of auditory sensitivity and degree of tonal differentiation are concerned, has been developed for a particular purpose—namely, the reception of high-frequency sounds within a limited range.

Echolocation in other mammals. Among the mammals possessing echolocation are the toothed whales. These animals probably produce sounds in the water in two ways: with the larynx and with the complex system of passages connected to the blowhole, which is a nostril in the top of the head. Although many different types of sounds are possible, during echolocation they consist mainly of a rapid series of clicks. These clicks contain many components, but the principal energy is in the high frequencies, from perhaps 50,000 to as much as 200,000 hertz. The use of such high frequencies by these animals is a requirement for effective echolocation in water. Because the velocity of sound is greater in water than in air, the wavelengths are longer; therefore, in order for echolocation to attain the same effectiveness of object discrimination as that achieved by the bat with aerial sounds, an aquatic animal has to use frequencies at least five times as high.

High-frequency clicks of toothed whales

Whales have good vision when submerged, and apparently their eyes remain fairly serviceable when their heads are out of water. Dolphins can be trained to strike targets or leap over obstacles held several feet above the surface of the water. For many tasks, however, they use echolocation very effectively, such as when catching fish at night or when visibility is poor in murky water. Dolphins have been trained to make fine discriminations of objects when their vision has been completely excluded by blindfolding. Echolocation of some form and degree of effectiveness is suspected in still other animals, such as shrews and sea lions, but the evidence is meagre thus far. (E.G.W.)

HUMAN SENSORY RECEPTION

Ancient philosophers called the human senses "the windows of the soul," and Aristotle enumerated at least five senses—sight, hearing, smell, taste, and touch—and his influence has been so enduring that many people still speak of the five senses as if there were no others. Yet, the human skin alone is now regarded as participating in (mediating) a number of different modalities or senses (*e.g.,* hot, cold, pressure, and pain). The modern sensory catalogue also includes a kinesthetic sense (sense organs in muscles, tendons, and joints) and a sense of balance or equilibrium (so-called vestibular organs of the inner ear stimulated by gravity and acceleration). In addition, there are receptors within the circulatory system that are sensitive to carbon dioxide gas in the blood or to changes in blood pressure; and there are receptors in the digestive tract that appear to mediate such experiences as hunger and thirst.

Not all receptors give rise to direct sensory awareness; circulatory (cardiovascular) receptors function largely in reflexes that adjust blood pressure or heart rate without the person being conscious of them. Though perceptible as hunger pangs, feelings of hunger are not exclusively mediated by the gastric (stomach) receptors. Some brain cells may also participate as "hunger" receptors. This is especially true of cells in the lower parts of the brain (such as the hypothalamus) where some cells have been found to be sensitive to changes in blood chemistry (water and other products of digestion) and even to changes in temperature within the brain itself.

General considerations of sensation

BASIC FEATURES OF SENSORY STRUCTURES

Classification by stimuli

One way to classify sensory structures is by the stimuli to which they normally respond; thus, there are photoreceptors (for light), mechanoreceptors (for distortion or bending), thermoreceptors (for heat), chemoreceptors (*e.g.*, for chemical odours), and nociceptors (for painful stimuli). This classification is useful because it makes clear that various sense organs can share common features in the way they convert (transduce) stimulus energy into nerve impulses. Thus, auditory cells and vestibular (balance) receptors in the ear and some receptors in the skin all respond similarly to mechanical displacement (distortion). Because many of the same principles apply to other animals, their receptors can be studied as models of the human senses. In addition, many animals are endowed with specialized receptors that permit them to detect stimuli that man cannot sense. A snake (the pit viper) boasts a receptor of exquisite sensitivity to "invisible" infrared light; some insects have receptors for ultraviolet light and for pheromones (chemical sex attractants and aphrodisiacs unique to their own species) thereby also exceeding human sensory capabilities.

Regardless of their specific anatomical form, all sense organs share basic features. (1) They contain receptor cells which are specifically sensitive to one class of stimulus energies, usually within a restricted range of intensity. Such selectivity means that each receptor can be said to have its own "adequate" or proper or normal stimulus, as, for example, light is the adequate stimulus for visual experience. Nevertheless, other energies ("inadequate" stimuli) can also activate the receptor if they are sufficiently intense. Thus, one may "see" pressure when, for example, the thumb is placed on a closed eye and one experiences a bright spot (phosphene) seen in the visual field at a position opposite the touched place. (2) The sensitive mechanism for each modality is often localized in the body at a receiving membrane or surface (such as the retina of the eye) where transducer neurons (sense cells) are to be found. Often the sensory organ incorporates accessory structures to guide the stimulating energy to the receptor cells; thus, the normally transparent cornea and lens within the eye focus light on the retinal sensory neurons. In some cases, blindness can be cured by surgically removing a lens that has grown opaque from cataract in order to permit light once again to reach the retina. Additional postoperative optical correction in the form of a contact lens or eyeglasses is necessary to compensate for the missing lens. Retinal nerve cells themselves are more or less shielded from nonvisual sources of energy by the surrounding structure of the eye; but mild electrical currents delivered to most sense organs, including the eye, can produce sensory experiences appropriate to the specific organ.

The generalized electrical nature of neural function largely accounts for the effectiveness of such currents in evoking a full range of different sensations. (3) The primary transducers or sensory cells in any receptor structure normally connect (synapse) with secondary, ingoing (afferent) nerve cells that carry the nerve impulse along. In some receptors, such as the skin, the individual primary cells possess threadlike structures (axons) that may be yards long, winding from just beneath the skin surface through subcutaneous tissues until they reach the spinal cord. Here each axon from the skin terminates and synapses with the next (second-order) neuron in the chain. By contrast, each primary receptor cell in the eye has a very short axon that is contained entirely in the retina, making synaptic contact with a network of several types of second-order (internuncial) cells, which, in turn, make synaptic contact with third-order neurons called bipolar cells, all still in the retina. The bipolar-cell axons extend afferently beyond the retina, leaving the eyeball to form the optic nerve, which enters the brain to make further synaptic connections. If this visual system is considered as a whole, the retina may be said to be an extended part of the brain on which light can directly fall. (4) From such afferent nerves, still higher order neurons make increasingly complex connections with anatomically separate pathways of the brainstem and deeper parts of the brain (*e.g.*, the thalamus) that eventually end in specific receiving areas in the cerebral cortex (the convoluted outer shell of the brain). Different sensory receiving areas are localized in particular regions of the cortex; *e.g.*, occipital lobes in back for vision, temporal lobes on the sides for hearing, and parietal lobes toward the top of the brain for tactual function.

Sensory parts of the brain

APPROACHES TO THE STUDY OF SENSING

The science of the human senses is truly interdisciplinary. Philosophers, physicians, anatomists, physical scientists, physiologists, psychologists, and others have all joined in studying sensory activities. Some of their earliest work was anatomical, an approach that continues to be fruitful. Physical scientists, particularly physicists and chemists, made especially important contributions to an understanding of the nature of stimulus energies (*e.g.*, acoustic, photic, thermal, mechanical, chemical); in the process, they also carried out many fundamental measurements of human sensory function. Hermann L.F. von Helmholtz, a 19th century German scientist who was a physicist, physiologist, and psychologist, studied the way in which sound waves and light are received (sensed or detected) and also how they are interpreted (perceived) by people.

Modern studies of sensation have been enhanced by contemporary devices permitting the precise production and control of sensory stimuli. With other kinds of instruments, physiologists have been able to probe the electrical signals generated by sensory cells and afferent nerve fibres to provide a biophysical analysis of sensory mechanisms. Psychophysics embraces the study of the inner (private, subjective) aspects of sensation in terms of outer (public, objective) stimulus energies. One of the oldest and most classical approaches to the study of sensation, psychophysics includes the study of people's reports of their sensations when they are stimulated: of their ability, for example, to match tones of equal loudness, to detect stimulus differences, and to estimate sensory magnitude or intensity under conditions of controlled stimulation. Psychophysical research continues as an active enterprise particularly among modern psychologists.

An old philosophical notion that "mind" is but a clean slate or tablet (*tabula rasa*) until written on by impressions from the senses no longer seems fully tenable; human infants, for example, show inborn (innate) ways of sensing or perceiving at birth. In its modern form, the problem of learned versus innate factors in sensory experience is studied in terms of the extent to which the genetically determined structure and function of sense organs and brain depend upon stimulation and experience for their proper maturation. Poverty of stimulation (sensory deprivation) in an infant's early life increasingly is being documented as detrimental to the full flowering of mature perceptual and intellectual functions. Since this sort of evidence may indeed lend some support to the notion of the *tabula rasa*, modern investigators give credence both to nativistic (based on heredity) and empiricistic (based on learning) interpretations of human sensory function (see also LEARNING AND COGNITION).

Emotional aspects of sensation

A distinction between the discriminatory (epicritic) and emotional (protopathic) features of sensations was made by Sir Henry Head (1861–1940), a British neurologist, who noted, for example, that after a sensory nerve from the skin had been cut, the first sensations to recover as the nerve healed appeared to be diffuse, poorly localized, and extremely unpleasant. He theorized that this initial lack of sharp discrimination associated with unpleasant experience reflected the properties of a primitive protopathic (emotional) neural system which regenerated first. He held that this system subserves pain and the extremes of temperature and pressure sensation usually associated with an affective (emotional) tone. Because recovery of fine tactual discrimination, sensitivity to lightly graded stimuli, and the ability to localize points touched on the skin returned later, Head posited the existence of another discriminatory system. While later research has not confirmed his theory, the sequence of changes in the recovery following nerve injury is most typical.

Chemical-visceral sensations particularly have hedonic

(pleasure–pain) properties. Most people tend to refer to odours and tastes as pleasant or unpleasant; thus, the chemical senses are closely tied to motivation, to preferences, or to aversions. Although reflex licking or sucking is stimulated by tactual stimulation of lips and mouth, newborn infants tend to suck longer and harder when the stimulus has clear hedonic value—*e.g.*, avidly turning their lips toward a nipple to get a sweet taste. (See also EMOTION, HUMAN.) Apparently, one's sweet tooth is largely nativistic, in that it requires little prior learning. The craving for salt (especially heightened under conditions of salt deprivation) likewise appears to be widespread throughout the animal kingdom without prior learning. The role of taste and smell as innate factors in behaviour may not be quite so influential in humans as in other animals. People's food habits and preferences are strongly overladen with custom and tradition; that is, they are learned in large measure.

In the modern era, the language of communication engineering has been found to be useful in describing human senses. Thus, each sensory modality may be described as a channel that receives stimulus information (input), that processes and stores the information, and that retrieves it as needed for the effective behaviour (output) of the individual. In addition, modern engineering has provided devices (*e.g.*, radio, television, radar, the electron microscope) that serve to extend the range and power of people's senses; in the last analysis, however, all such devices convert (transduce) their information back to a form of stimulus energy that is directly perceptible to the unaided senses. Thus a television set is a transducer that converts imperceptible electromagnetic waves into visual and auditory signals. For some special purposes, people may employ alternative sensory channels, as when blind people use Braille or other tactual input as substitutes for the missing visual channels. While the chemical senses have little function in symbolic communication among people, the use of perfumes in romantic signalling is a notable exception. In general, however, the human chemical senses are more directly involved in physiological survival—*e.g.*, warning that a putrid fish is dangerous to eat. One's physical well-being also rests heavily on proprioceptors (for sensing one's own bodily position) and on the sense of balance. These structures, monitoring (feeding back information on) one's bodily orientation in space, provide crucial sensory feedback for guiding one's movements (see also PERCEPTION, HUMAN).

Survey of some of the human senses

CUTANEOUS (SKIN) SENSES

It was observed above that studies of cutaneous sensitivity yield evidence that the human senses number more than five. There is evidence for two pressure senses (for light and for deep stimulation), for two kinds of temperature sensitivity (warm and cold), and for a pain sense. In the 1880s, experimental findings that the human skin is punctate (selectively sensitive at different points) gave clear indication of a dissociation among functions once lumped together as the sense of touch. Mapping the skin with a fine bristle or with a narrow-tipped (warm or cold) cylinder showed that there were different spots of maximum sensitivity to pressure, warm, and cold. When stimulated between the spots on the skin, people reported no such sensations. Pain spots also can be located with a finely pointed needle, but the punctate character is less striking since pain seems to be widespread when stimulus intensity is increased. The number of spots is greatest for pain, next for touch, then for cold, and least for warm. Efforts to identify specific receptor cells for each of these sensitive points have been the subject of much debate and still pose a problem that is not completely settled.

Mapping sensitive spots of skin

Nerve function. Microscopic examination of the skin reveals a variety of nerve terminals; there are free nerve endings (which are most common), so-called Ruffini endings, and encapsulated endings, such as the Pacinian corpuscle, Meissner's corpuscle, or Krause end bulb, all named for investigators who discovered them (Figure 32). At one time it was thought that each of these specialized

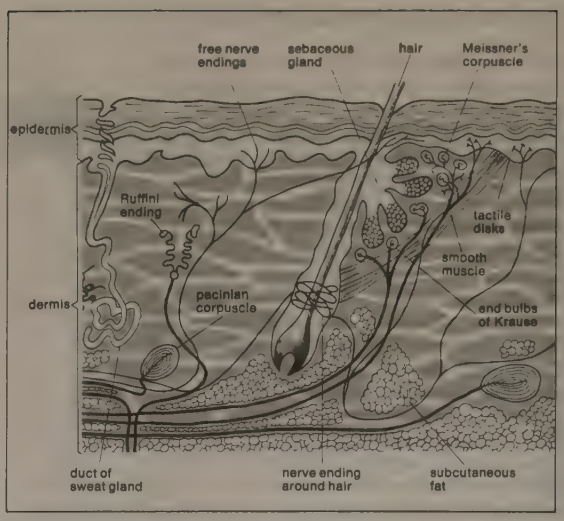


Figure 32: Microscopic view of human skin.

From E. Gardner, *Fundamentals of Neurology*, 5th ed. (1968), W.B. Saunders Company, Philadelphia

structures mediated one of the cutaneous modalities, but efforts to extirpate (surgically remove) the nerve endings under the spots have yielded only questionable data. Further, the cornea of the eye shows only free nerve endings; yet pain, pressure, and some temperature sensations can be elicited by stimulating the corneal surface.

Electrical recordings from the cutaneous nerves of laboratory animals suggest a much wider variety of receptors than are encompassed by the reports people give of their sensations. Some nerve endings seem to respond only to one type of stimulus (*e.g.*, to pressure stimuli of very light weight or to slight temperature changes); others exhibit a broad range of sensitivity. Some receptors show combined sensitivity to both temperature and pressure. In some cases only special types of mechanical stimulation (such as rubbing) may be effective. Furthermore, there is extensive overlap in the areas of skin (receptor fields) for the individual nerve fibres examined, suggesting a neural integration of overlapping afferent inputs of skin nerves. A model of the sensory system that envisages only a single nerve fibre serving one tactual spot is clearly contradicted.

On the other hand, some tactual receptors (*e.g.*, Pacinian corpuscles) respond only to mechanical deformation. This corpuscle is an onion-shaped structure of non-neural (connective) tissue built up around the nerve ending; indeed, the distinctive corpuscle, if anything, reduces the mechanical sensitivity of the nerve terminal itself. If the onion-like capsule is entirely removed, mechanical sensitivity not only remains but is somewhat greater than when the capsule is present.

Variations in sensory nerve fibres

In addition to the differences in the sensory end structures of the skin, the afferent nerve fibres (axons) from them also show diversity. The nerve fibres range in size from large myelinated (sheathed) axons of 10 to 15 microns diameter down to extremely small unmyelinated fibres measuring only tenths of microns across. Fatter axons tend to conduct nerve impulses more rapidly than do small fibres; when axons of different diameters form a single bundle (a nerve), they constitute a so-called mixed nerve. Thus, electrical records from a mixed nerve show what are labelled A (fast), B (medium), and C (slow) components that reflect the typical speeds at which axons of different diameters conduct. Although such specialized capsules as Pacinian corpuscles tend to be associated with larger diameter axons, and temperature-sensitive endings tend to be associated with medium-size fibres, a unique relation of each of the skin modalities with one of the A, B, or C fibre groups cannot be supported. All of the cutaneous senses seem to be associated with some fibres of all diameters; furthermore, the C fibres (once thought to be restricted to the pain function) display quite specific sensitivities to nonpainful stimuli applied to the skin.

A major neural pathway for tactual impulses runs along the back (in the dorsal columns) of the spinal cord. Af-

ferent fibres enter the cord from the cutaneous nerves and ascend without synaptic break in one (the ipsilateral) dorsal column. This is a very rapidly conducting pathway shared by fibres that mediate sensations of deep pressure and also kinesthesis. Other tactual, temperature, and pain information crosses the spinal cord close to the level of entry of the sensory fibres and ascends to the brain in contralateral pathways of the cord (the lateral and ventral spinothalamic tracts).

Each of the nerves distributed along the spinal cord contains a sensory bundle that serves a well-defined strip of skin (a dermatome) about 2.5 centimetres (one inch) or more wide on the body surface. Successive spinal nerves overlap, so that each place on the skin represents two and sometimes three dermatomes; this yields a segmented pattern of strips over the body from head to toe. All dermatomes feed into a single relay centre (the sensory thalamus) deep within the brain, where a precise three-dimensional layout of tactual sensitivity at the body surface can be found. The neurons in this part of the thalamus (the ventral posterolateral nucleus) are specific to particular skin senses (such as pressure) and form small and precise receptor fields. There is a second more diffuse thalamic system (in the posterior thalamic nuclei) where the receptor fields are large, perhaps bilateral, on the left and right sides, perhaps including one whole side of the body. The receptor fields here or the types of stimuli to which they respond are not clearly delineated.

The dissociation of cutaneous senses is dramatically demonstrated in the course of some diseases; for example, in a disorder called syringomyelia, degeneration of the central canal of the spinal cord leads to loss of pain and temperature sensitivity. Nevertheless, the sufferer still can experience pressure. In some instances there may be a complete absence of pain sensitivity with disastrous consequences for the welfare of the afflicted person. Such individuals are bruised and cut and even lose parts of the body because they are unable to sense the dangerous (painful) characteristics of stimuli. Among people born with total absence of the pain sense, there may not be demonstrable anatomic abnormality. Still other instances of dissociation of pain versus pressure occur in surgical procedures (such as tractotomy) in which spinal tracts or parts of the nerves leading into the brainstem are selectively cut. Such operations are designed specifically to relieve pain without unduly diminishing pressure sensitivity.

Pathways from the specific (ventral posterolateral) thalamus end (or project) in a narrow band of brain cortex (the posterior rolandic cortical sensory area in man) where there is a point-for-point representation of the body surface on the cortical surface. The cortical projection of the more diffuse (posterior) thalamic system is less well charted. There thus appears to be a dissociation between those tactual structures that are highly specific and those that are more generalized.

Tactual psychophysics. The mixture of sensitivities within a given patch of skin provides a ready basis for the concept of adequate stimulation. Sometimes, for example, a cold spot responds to a very warm stimulus, and the person experiences what is called paradoxical cold. The sensation of heat from a hot stimulus presumably arises from the adequate stimulation of warmth receptors combined with the inadequate or inappropriate (albeit effective) stimulation of cold and pain receptors.

Paradoxical cold

Human ability to barely detect pressure (*i.e.*, the human pressure threshold) generally appears when a tension of about 0.85 grams per square millimetre (equivalent to about 1.2 pounds per square inch) of skin surface is applied on the back of the hand. Thus a force of 85 milligrams applied to a stimulus hair (or bristle) of 0.1 square millimetre is just about enough to elicit the experience of pressure. The energy of impact at pressure threshold is very much greater than that required for hearing or seeing, the skin requiring on the order of 100,000,000 times more energy than the ear and 10,000,000,000 times more energy than the eye. Differential pressure discrimination (the ability to detect just noticeable differences in intensity) requires changes of roughly 14 percent at maximum sensitivity.

Adaptation to pressure is well known; one's awareness of a steadily applied bristle fades and ultimately disappears. As a result people are rarely aware of the steady pressure of their clothing unless movement brings about a change in stimulation. Most dramatic and perhaps best known among tactual experiences is adaptation to thermal stimulation. Continued presentation of a warm or cold stimulus leads to reduction or disappearance of the initial sensation and an increase in threshold values. Total obliteration of thermal sensation through adaptation occurs in the range from about 16° to 42° C (61° to 108° F). If one hand is placed in a bowl of hot (40° C [104° F]) water and adapted to that, and at the same time the other hand is adapted to cold (20° C [68° F]) water, then when both hands are simultaneously placed in lukewarm (30° C [86° F]) water, the previously cooled hand feels warm and the other hand feels cold. This effect was once interpreted as evidence for a single temperature sense, but careful study shows that there are indeed two kinds of temperature receptors, both of which show adaptation. Cold receptors are characterized by an electrical discharge on sudden cooling, normally showing no response to sudden warming; similarly appropriate electrical responses are made by warmth receptors. Both receptors show steady discharges selectively depending on temperature; maximum discharge typically occurs between 38° to 43° C (100° to 109° F) for individual warmth cells and between 15° and 34° C (59° and 93° F) for cold receptors. These temperature receptors show no electrical response to weak mechanical stimulation in either laboratory animals or human subjects.

Pain

Pain is least understood among all the human senses. The pattern of stimulation is more crucial in pain than in any other sense. A single brief electric shock to the skin or to an exposed nerve may not elicit the experience of pain; yet it tends to become painful upon repetitive stimulation. Cutaneous pain is often sensed more sharply than is pain associated with deep tissues of the body (*e.g.*, viscera). Certain areas of the body are relatively analgesic (free of pain); for example, the mucous lining of the cheek into which one can bite shallowly without discomfort. The organs of the abdominal cavity are usually insensitive to cutting or burning, but traction or stretching of hollow viscera is painful (as when the stomach is distended by gas). Pain displays sensory adaptation, although the process appears to be more complex than it is for other sensory modalities. Thus, the intensities of headaches, toothaches, and pains from injury often show cyclic fluctuations, possibly from such factors as changes in blood circulation or in degree of inflammation. The visceral pains, those of dental origin, or of diseased tissues can be reduced by analgesic drugs, which tend to be less effective on cutaneous pain. Pain has a strong emotional context. In certain cases, after frontal lobotomies (a type of brain surgery) have been performed, a person may report that he still feels the pain of a pin prick or other irritation but that he does not find it as disturbing or emotionally disruptive as he did before the lobotomy. Many phenomena indicate the powerful role of the brain and spinal cord in sensing potentially painful sensory input. Indeed, according to one theory, a so-called gate control system in the spinal cord modulates (increases or decreases) sensory input from the skin to determine whether the input is perceived as painful. This theoretical formulation also may account for moment-to-moment fluctuations in the intensity of perceived pain despite the absence of any stimulus change. Such brain-mediated factors as emotional tension or past psychological experience are held to influence pain perception by acting upon this spinal gate control system.

Itching

Itching seems to bear the same relation to pain as tickle does to pressure. The experience usually lasts long enough to demand attention and (like tickle) normally leads to a response such as rubbing or scratching the affected area. A number of skin disorders are accompanied by itching, presumably from a fairly low level of irritation in the affected area (which also may be produced in undiseased skin). While a single shock by a low-intensity electrical spark normally produces no sensation, a repetitive pattern of such shocks may induce an itch not unlike that produced by an insect bite. Itching also may occur as an aftereffect of the sharp pricking sensation produced by single strong

shocks, presumably because the nerves continue to produce a patterned afterdischarge following the cessation of the stimulus. Nonpainful tactile pattern stimulation is exemplified by vibration. Different frequencies of vibration are readily discriminated and a tactual communication system employing vibrations on the skin has been devised, particularly for people who cannot see or hear. Further research will probably reveal other ways of utilizing the fine discriminatory capacities of the cutaneous senses as substitutes for other sensory avenues of communication.

KINESTHETIC (MOTION) SENSE

It is a common experience that, with the eyes closed, one is aware of the positions of his legs and arms and can perceive the active or passive movement of a limb and its direction. The term kinesthesis (literally "feeling of motion") has been coined for this sensibility.

Nerve function. Four types of sensory structures are widely distributed in muscles, tendons, and joints (Figure 33): (1) the neuromuscular spindle consists of small, fine

From E. Gardner, *Fundamentals of Neurology*, 5th ed. (1968) W.B. Saunders Company, Philadelphia

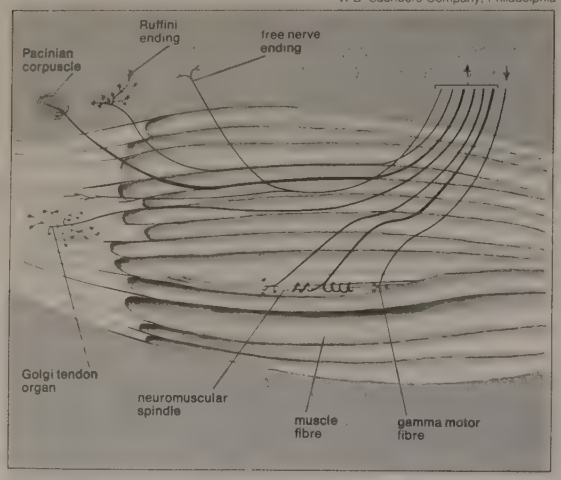


Figure 33: Sensory structures of muscle, tendon, and joint. Arrows indicate direction of conduction. (Motor nerves and motor nerve endings on contractile muscle fibres not shown.)

muscle fibres around which sensory fibre endings are wrapped; (2) the Golgi tendon organ consists of sensory nerve fibres that terminate in a rich branching encapsulated within the tendon; (3) joint receptors (as in the knee) consist of "spray-type" endings (Ruffini) and Golgi-type and Pacinian corpuscles within the joints; and (4) free nerve endings. All these receptors combine to provide information on active contraction, passive stretch of muscle fibres, and tension whether produced actively or passively. In passive stretch both the muscle-spindle receptors and the tendon receptors send trains of impulses over their sensory (afferent) nerves; in active contraction the spindles show a silent period of neural activity when tension on the parallel fibres is unloaded, while the tendon receptors discharge just as when stretch is passive.

The muscle spindle is contractile in response to its own small-diameter, so-called gamma motor (efferent) fibre. The receptors and the gamma fibres of the muscle spindle form a neuromuscular loop that serves to insure that tension on the spindle is maintained within its efficient operating limits. The excitability of the muscle spindle also can be influenced through other neural pathways that control the general level of excitability of the central nervous system (brain and spinal cord). Activity of the descending reticular formation (a network of cells in the brainstem) may enhance the contraction of the spindle and hence influence its neural discharges.

The knee jerk

Muscle and tendon receptors combine to play an intimate and crucial role in the regulation of reflex and voluntary movement. Much of this control is automatic (involuntary) and not directly perceptible except in the aftereffects of movement or change of position. The knee jerk that follows a tap just below the kneecap of a freely hanging leg is one such involuntary reflex. Sensory (afferent) impulses from stretching the receptors (*e.g.*, in the muscles) relay to the spinal cord and activate a path to the motor (efferent) nerves leading back to the same muscle. The knee jerk is a purely spinal reflex response (the brain is not required) which is tested usually to determine nerve damage or other interference with the spinal-cord motor mechanisms. Besides producing loss of knee jerk, a disease such as syphilis may lead to locomotor ataxia (a clumsy and stumbling gait) when the germ (called a spirochete) attacks the sensory nerves of the cord's dorsal column. The result is that the sufferer has difficulty sensing the position of his limbs. Another general function of the muscle receptors is the maintenance of muscle tone (or tonus: partial contraction) to permit rapid response (fast reaction time) to stimulation. In normal conditions the muscle has tone and is ready to go; but, when it is without motor stimulation (deafferented), the muscle is flaccid, showing little tone. One's upright posture depends on the tonus of opposing (extensor and flexor) muscles in response to the effects of gravity.

The exact contribution of the muscle receptors to sensation is not entirely understood. It seems clear, however, that they are not essential to the sensation of bodily position. It has been suggested that the appreciation of passive movement of the limbs probably comes largely from the joints, since, after anesthetizing the overlying skin and muscles (*e.g.*, with cocaine), sensibility to the limb movement seems little affected. Evidence also shows that very few of the impulses arising from the muscle receptors themselves reach the cerebral cortex; instead, they ascend in the spinal pathways to another part of the brain (the cerebellum) where they interact in the automatic control of bodily movement. Impulses arising from the joint receptors, on the other hand, have been recorded in both the thalamus and brain cortex, the degree of angular displacement of a joint being reflected systematically in these structures by the frequency of nerve impulses. Symptoms of some diseases also emphasize the importance of joint sensitivity. When bone disease, for example, destroys only the joint receptors the person's ability to appreciate posture and movement is lost.

The feedback system. The feedback system leading to muscle tonus is a most delicately balanced mechanism. The gamma loop feeds back information that serves to maintain the muscle tonus and postural adjustments appropriate to the efficient performance of different voluntary actions. The afferent input from the muscle spindles via the spinal cord to the brain's cerebellum traverses an extensive circuit involving interactions of excitatory and inhibitory processes, the end result of which helps insure smooth and finely coordinated movements. Disease or other neurological damage in the cerebellum is characterized by distortions of movement and posture. Some sufferers of cerebellar disorders display crude, overactive motor activity (ballistic movement) that overshoots the mark in attempted voluntary coordination. Other victims of cerebellar disease display what resembles a drunken gait, jerkily stumbling and swaying along. By relying on other sensory cues (*e.g.*, visual and tactual), some people are able to compensate for awkward and uncoordinated movements associated with cerebellar damage.

Cerebellar disorders

VESTIBULAR SENSE (EQUILIBRIUM)

Function of the inner ear. The human inner ear contains parts (the nonauditory labyrinth or vestibular organ) that are sensitive to acceleration in space, rotation, and orientation in the gravitational field. Rotation is signalled by way of the semicircular canals, three bony tubes in each ear that lie embedded in the skull roughly at right angles to each other. These canals are filled with fluid (endolymph); in the ampulla of each canal are receptor cells with fine hairs that project up into the fluid to be displaced as the endolymph lags behind when rotation begins. When rotation is maintained at a steady velocity, the fluid catches up, and stimulation of the hair cells no longer occurs until rotation suddenly stops, again circulating the endolymph. Whenever the hair cells are thus stimulated, one normally experiences a sensation of rotation in space. During rotation one exhibits reflex nystagmus (back-and-

forth movement) of the eyes. Slow displacement of the eye occurs against the direction of rotation and serves to maintain the gaze at a fixed point in space; this is followed by a quick return to the initial eye position in the direction of the rotation. Stimulation of the hair cells in the absence of actual rotation tends to produce an apparent swimming of the visual field, often associated with dizziness and nausea; compensatory postural adjustments commonly follow.

Two sacs or enlargements of the vestibule (the saccule and utricle) react to steady (static) pressures; *e.g.*, those of gravitational forces. Hair cells within these structures are covered by a gelatinous cap in which are embedded small granular particles of calcium carbonate (otoliths) that weigh against the hairs. When iron particles are implanted in the same structures of a fish, they may be displaced by externally applied magnets. In this way the fish can be made to assume inverted or other unusual positions in the water. In man, unusual stimulation of the vestibular receptors and semicircular canals also can give rise to sensory distortions in visual and motor activity. The resulting discord between one's visual and motor responses and the external space (as aboard ship in a heaving sea) often leads to nausea and disorientation (*e.g.*, seasickness). In space flight these sensory systems usually are not stimulated except as the weightless astronaut affects them by his own movements. Such abnormal gravitational and acceleratory forces apparently contribute to the nausea or disequilibrium sometimes reported by people in outer space. Training before space flight reduces the likelihood and severity of these symptoms.

Seasickness

Factors affecting equilibrium. In some diseases (*e.g.*, ear infections), irritation of vestibular nerve endings occurs, and the sufferer may be subject to falling as well as to spells of disorientation and vertigo (dizzy confusion). Similar symptoms may be induced by flushing hot and cold water into the outer opening of the ear, since the temperature changes produce currents in the endolymph of the semicircular canals. This caloric (temperature) effect is used in clinical tests for vestibular functions and in physiological experiments. Externally applied electrical currents may also stimulate the nerve endings of the vestibule. When current is applied to the right mastoid bone (just behind the ear), nystagmus to the right tends to occur, associated with a reflex right movement of the head; movement tends to the left for the opposite mastoid. In man, destruction of the labyrinth in only one ear causes vertigo and other vestibular symptoms, such as nystagmus, inaccurate pointing, and tendency to fall. (The vestibular functions of the ear are described in detail below; see *Human hearing*, with special attention to the section *Ear diseases and hearing disorders: the inner ear.*)

TASTE (GUSTATORY) SENSE

The sensory structures for taste in man are the taste buds, clusters of cells contained in goblet-shaped structures (papillae) that open by a small pore to the mouth cavity. A single bud contains about 50 to 75 slender cells, all arranged in a bananalike cluster pointed toward the gustatory pore (see also *Chemoreception* above). These are the taste receptor cells, which differentiate from the surrounding epithelium, grow to mature form, and then die out to be replaced by new cells in a turnover period as short as seven to 10 days. The various types of cells in the taste bud appear to be different stages in this turnover process. Slender nerve fibres entwine among and make contact usually with many cells. In man and other mammals, taste buds are located primarily in fungiform (mushroom-shaped), foliate, and circumvallate (walled-around) papillae of the tongue or in adjacent structures of the palate and throat. Many gustatory receptors in small papillae on the soft palate and back roof of the mouth in human adults are particularly sensitive to sour and bitter, whereas the tongue receptors are relatively more sensitive to sweet and salt. Some loss of taste sensitivity suffered among wearers of false teeth may be traceable to mechanical interference of the denture with taste receptors on the roof of the mouth.

Locations of taste buds

Nerve supply. There is no single sensory nerve for taste in vertebrates. In man, the anterior (front) two-thirds of the tongue is supplied by one nerve (the lingual nerve), the back of the tongue by another (the glossopharyngeal nerve), and the throat and larynx by certain branches of a third (the vagus nerve), all of which subserve touch, temperature, and pain sensitivity in the tongue as well as taste. The gustatory fibres of the anterior tongue leave the lingual nerve to form a slender nerve (the chorda tympani) that traverses the eardrum on the way to the brain stem. When the chorda tympani at one ear is cut or damaged (by injury to the eardrum), taste buds begin to disappear and gustatory sensitivity is lost on the anterior two-thirds of the tongue on the same side. Impulses have been recorded from the human chorda tympani, and good correlations have been found between the reports people give of their sensations of taste and of the occurrence of the afferent nerve discharge. The taste fibres from all the sensory nerves from the mouth come together in the brainstem (medulla oblongata). Here and at all levels of the brain, gustatory fibres run in distinct and separate pathways, lying close to the pathways for other modalities from the tongue and mouth cavity. From the brain's medulla, the gustatory fibres ascend by a pathway to a small cluster of cells in the thalamus and thence to a taste-receiving area in the anterior cerebral cortex.

Physiological basis of taste. No simple relation has been found between chemical composition of stimuli and the quality of gustatory experience except in the case of acids. The taste qualities of inorganic salts (such as potassium bromide, a sedative) are complex; epsom salt (magnesium sulfate) commonly is sensed as bitter, while table salt (sodium chloride) is typical of sodium salts, which usually yield the familiar saline taste. Experiences of sweet and bitter are elicited by many different classes of chemical compound.

Types of taste receptors

Theorists of taste sensitivity classically posited only four basic or primary types of human taste receptors, one for each gustatory quality: salty, sour, bitter, and sweet. Yet, recordings of sensory impulses in the taste nerves of laboratory animals show that many individual nerve fibres from the tongue are of mixed sensitivity, responding to more than one of the basic taste stimuli, such as acid plus salt or acid plus salt plus sugar. Other individual nerve fibres respond to stimuli of only one basic gustatory quality. Most numerous, however, are taste fibres subserving two basic taste sensitivities; those subserving one or three qualities are about equal in number and next most frequent; fibres that respond to all four primary stimuli are least common. Mixed sensitivity may be only partly attributed to multiple branches of taste nerve endings. In man, experiences of sugars, synthetic sweeteners, weak salt solutions, and the taste of some unpleasant medicines are blocked by a drug (gymnemic acid) obtained from *Gymnema* bushes native to India. Among some laboratory animals, gymnemic acid blocks only the nerve response to sugar, even if the fibre mediates other taste qualities. Such a multiresponsive fibre still can transmit taste impulses (*e.g.*, for salt or sour), so that blockage by the drug can be attributed to chemically specific sites or cells in the taste bud.

In some species of animals (*e.g.*, the cat), specific taste receptors appear to be activated by water; these so-called water receptors are inhibited by weak saline solutions. Water taste might be considered a fifth gustatory quality in addition to the basic four.

The qualities of taste. *Sour.* The hydrogen ions of acids (*e.g.*, hydrochloric acid, HCl) are largely responsible for the sour taste; but, although a stimulus grows more sour as its hydrogen ion (H^+) concentration increases, this factor alone does not determine sourness. Weak organic acids (*e.g.*, the acetic acid in vinegar) taste more sour than would be predicted from their hydrogen ion concentration alone; apparently the rest of the acid molecule affects the efficiency with which hydrogen ions stimulate.

Salt. Although the salty taste is often associated with water-soluble salts, most such compounds (except sodium chloride) have complex tastes such as bitter-salt or sour-salt. Salts of low molecular weight are predominantly salty, while those of higher molecular weight tend to be bitter. The salts of heavy metals such as mercury have a metallic

taste, although some of the salts of lead (especially lead acetate) and beryllium are sweet. Both parts of the molecule (*e.g.*, lead and acetate) contribute to taste quality and to stimulating efficiency. In man the following series for degree of saltiness, in decreasing order, is found: ammonium (most salty), potassium, calcium, sodium, lithium, and magnesium salts (least salty). The order appears to vary for other animals.

Sweet. Except for some salts of lead or beryllium, the sweet taste is associated largely with organic compounds (such as alcohols, glycols, sugars, and sugar derivatives). Human sensitivity to synthetic sweeteners (*e.g.*, saccharin) is especially remarkable; the taste of saccharin can be detected in a dilution 700 times weaker than that required for cane sugar. The stereochemical (spatial) arrangement of atoms within a molecule may affect its taste; thus, slight changes within a sweet molecule will make it bitter or tasteless (Figure 34). Several theorists have proposed

From R W Moncrieff, *Chemical Senses* (1944), Leonard Hill Books

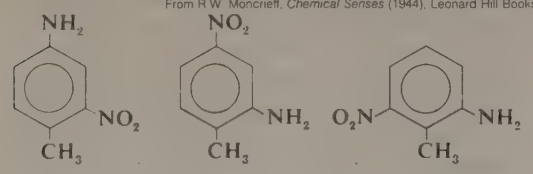


Figure 34: Effects of molecular arrangement on taste sensation (see text).

that the common feature for all of the sweet stimuli is the presence in the molecule of a so-called proton acceptor, such as the OH (hydroxyl) components of carbohydrates (*e.g.*, sugars) and many other sweet-tasting compounds. It has also been theorized that such molecules will not taste sweet unless they are of appropriate size.

Bitter. The experience of a bitter taste is elicited by many classes of chemical compounds and often is found in association with sweet and other gustatory qualities. Among the best known bitter substances are such alkaloids (often toxic) as quinine, caffeine, and strychnine. Most of these substances have extremely low taste thresholds and are detectable in very weak concentrations. The size of such molecules is theoretically held to account for whether or not they will taste bitter. An increase in molecular weight of inorganic salts or an increase in length of chains of carbon atoms in organic molecules tends to be associated with increased bitterness.

Taste blindness

A substantial minority of people exhibit specific taste blindness, an inability to detect as bitter such chemicals as PTC (phenylthiocarbamide). Taste blindness for PTC and other carbamides appears to be hereditary (as a recessive trait), occurring least frequently among American Indians and Africans, in about a third of Europeans, and in roughly 40 percent of the people in Western India. Evidence for such taste blindness among other animals has been observed only in anthropoid apes; apparently it appeared at a relatively late stage of evolution. Taste blindness for carbamides is not correlated with insensitivity to other bitter stimuli, the reasons for this being poorly understood.

Factors affecting taste sensitivity. Fluids of extreme temperature, especially those that are cold, may produce temporary taste insensitivity. People generally seem to taste most acutely when the stimulus is at or slightly below body temperature. When the tongue and mouth are first adapted to the temperature of a taste solution, sugar sensitivity increases with temperature rise, salt and quinine sensitivity decrease, and acid sensitivity is relatively unchanged. Gustatory adaptation (partial or complete disappearance of taste sensitivity) may occur if a solution is held in the mouth for a period of time. The effect of one adapting stimulus on the person's sensitivity for another one (cross adaptation) is especially common with substances that are chemically similar and that elicit the same taste quality. Adaptation to sodium chloride will reduce one's ability to sense the saltiness of a variety of the inorganic salts but will leave undiminished or even enhance such qualities as bitterness, sweetness, or sourness that were part of the taste of the salt before adaptation. Likewise, adaptation by one acid may reduce sensitivity to the sourness of other acids.

Contrast effects

Adaptation studies often are complicated by so-called contrast effects; for example, people say that distilled water tastes sweet following their exposure to a weak acid. Water may take on other taste qualities as well; following one's adaptation to a sour-bitter chemical (urea), water may taste salty. Adaptation tends to diminish or enhance the effect of a subsequent stimulus depending on whether the two stimuli normally elicit the same or a contrasting taste. Thus the adapted sweetness of water and all normally sweet-tasting substances are enhanced after one has tasted acid (sour). The bitterness of tea and coffee or the sourness of lemon are masked or suppressed by sugar or saccharin.

The human gustatory difference threshold (for a just noticeable difference in intensity) is approximately a 20 percent change in concentration. For very weak taste stimuli, however, the threshold sensitivity is poorer.

Food choice. One's ability to taste is intimately concerned in his eating habits or in his rejection of noxious substances. One of the earliest reflex responses of the infant, that of sucking, can be controlled by gustatory stimuli. Sweet solutions are sucked more readily than is plain water; bitter, salty, or sour stimuli tend to stop the sucking reflex.

Among insects, a very specific feeding reaction (proboscis extension) is so automatic that it is widely used as an index of taste stimulation. If a common housefly is held relatively immobile in wax, different parts of the mouth, legs, and body readily may be stimulated by a drop of solution. Sugar solution will make the fly extend its proboscis when a drop is applied to the legs or mouth parts. A fly that has been starved will show a positive response to a weak sugar solution that ordinarily would not affect one that is satiated. Addition of salt or acid to the sugar solution inhibits this response.

Many animals provide clear examples of beneficially selective feeding behaviour. Laboratory rats, when given unhampered choice of carbohydrates, proteins, vitamins, and minerals (each in a separate container), show consistent patterns of self-selection that may be modified by physiological stresses and strains. A rat made salt-deficient by removal of its adrenal glands, for example, will increase its intake of sodium chloride sufficiently to maintain health and growth; normally, such gland removal is fatal in the absence of salt-replacement therapy. Histories of similar effects have been reported in human beings, one dramatic case being that of a child with adrenal disorder who kept himself alive by satisfying an intense salt craving.

Among human adults, past experience shows a strong influence on eating habits, sometimes to the point that physiological well-being suffers. Food habits and other factors play a significant role in eating behaviour.

Poisonous substances often are unpalatable, but not invariably. Lead acetate, sometimes called sugar of lead, once was used as a sweetening agent with disastrous results before its potentially fatal effects were appreciated. Many palatable substances, including some synthetic sweeteners, are toxic; taste alone is not a reliable guide to safety. A rat poison, alpha-naphthylthiourea (ANTU), was developed in a relatively insoluble and therefore tasteless form; soluble forms of ANTU had all been rejected by the animals. Taste aversions also may be readily established by conditioning, even for substances that have been normally preferred. In one study, a rat tasted saccharin solution three hours before being exposed to enough radiation to become sick. When the animal recovered, it was found to have a strong aversion to the taste of saccharin. Other aversions selectively can be produced by injecting the individual with a nauseating drug following specific taste experience. An unusual finding is that long delays of up to several hours in the time between the presentation of the taste stimulus and the induction of illness do not prevent the conditioning. In most other studies, only brief intervals (perhaps up to minutes in duration) have been found to result in successful conditioning. Bait shyness developed by wild rats that survive poisoning strongly suggests conditioned

taste aversion. Positive preferences also are subject to conditioning, as when the tastes of drugs or vitamins become associated with the feelings of well-being they generate.

SMELL (OLFACTORY) SENSE

Olfactory receptors

In mammals the olfactory receptors are located high in the nasal cavity. The yellow-pigmented olfactory membrane in humans covers about 2.5 square centimetres (0.4 square inch) on each side of the inner nose. The olfactory sense receptor is a long thin cell ending in several delicate hairs (cilia) that project into and through the mucus that normally covers the nasal epithelium or lining (Figure 35).

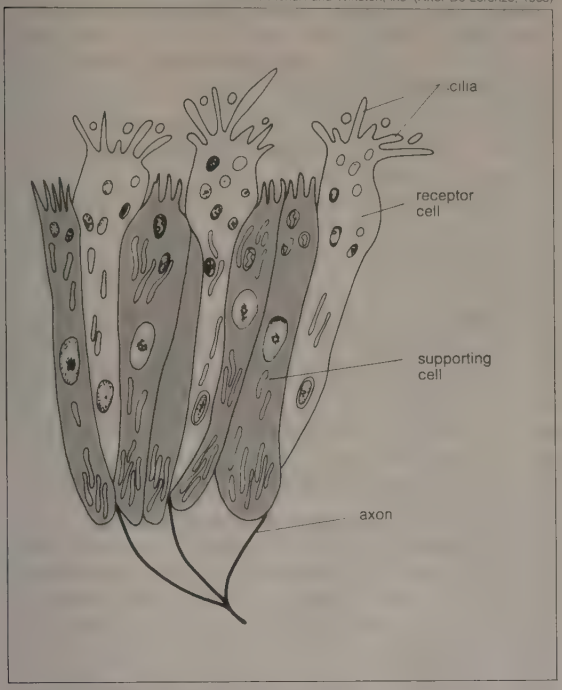


Figure 35: Olfactory lining of the nose.

Electron microscope photographs show from six to 12 olfactory cilia per cell. The end of each receptor narrows to a fine nerve fibre, which, along with many others, enters the olfactory bulb of the brain through a fine channel in the bony roof of the nasal cavity. The olfactory membrane of a young rabbit contains about 100,000,000 ciliated receptor cells that provide as much area as total skin surface of the animal.

Pain endings of the trigeminal nerve fibres are widely distributed throughout the human nasal cavity, including the olfactory region. Relatively mild odorants, such as orange oil, as well as the more obvious irritants, such as ammonia, stimulate such free nerve endings as well as the olfactory receptors.

Odorous molecules may be carried to the olfactory region by slight eddies in the air during quiet breathing, but vigorous sniffing brings a surge into the olfactory region. Odour sensitivity may be impaired by blocking the nasal passages mechanically, as when membranes are congested by infection.

It is generally agreed that the antennae of insects are their principal olfactory sites, but specific mouth parts may also bear endings for smell. The nonliving outer covering (cuticle) of such structures has been shown to be pierced by many ultramicroscopic holes through which odorous molecules enter, presumably to dissolve the fluid underneath. In vertebrates (such as man), olfactory stimulation occurs only after the odour molecule is dissolved in the mucus that covers the olfactory membrane. In spite of the wide biological gap between them, odour sensitivity of man and that of insect display certain similarities of mechanism.

Olfactory bulbs

In vertebrates the olfactory nerve fibres enter either of two specialized structures (olfactory bulbs), stemlike projections under the front part of the brain, to end in a microscopic series of intricate basket-like clusters called glomeruli. Each glomerulus receives impulses from about 26,000 receptors and sends them on through other cells, eventually to reach higher olfactory centres at the base of the brain. There are also fibres that cross over from one olfactory bulb to the other. When the olfactory bulbs are removed by surgery, the individual's ability to discriminate odours is lost; details of higher brain centres for smell are still unclear.

Olfactory qualities. The vocabulary of odour is rich with the names of substances that elicit a great variety of olfactory qualities. One of the best known published psychological attempts at classification was in 1916 on the basis of tests of more than 400 different scents on human observers. On the basis of the apparent similarities of perceived odour quality or confusions in naming, it was concluded that there were six main odour qualities: fruity, flowery, resinous, spicy, foul, and burned.

Electrical activity can be detected readily with fine insulated wires inserted into the olfactory bulb. Those portions of the bulb toward the anterior or oral (mouth) region in the rabbit are found to be more sensitive to water-soluble substances, whereas the more posterior or aboral (away from the mouth) parts of the olfactory bulb are more sensitive to fat-soluble substances. In addition, when very fine electrodes are used, individual cells (so-called mitral cells) are found to be sensitive to different groups of chemicals. Evidence for the existence of only a few primary receptors, however, does not emerge from such studies; a variety of different combinations of sensitivity has been found. Similarly, recordings from the primary receptor nerve fibres reveal different patterns of sensitivity. Electrical recording of this type also shows that olfactory sensitivity can be enhanced by a painful stimulus, such as a pinch on the foot of an experimental animal. This appears to be a reflex that serves to enhance the detection of dangerous stimuli in the environment. Different parts of the olfactory neural pathways seem to be selectively tuned to discriminate different classes of olfactory information. Thus, the third- and fourth-order olfactory neurons found beyond the olfactory bulb of the rat seem particularly concerned with distinguishing the odour of sexually receptive females. These neurons appear to be especially important in the preference the male rat shows for the smell of urine from a female in heat.

Characteristics that impart odour

Odorous substances. To be odorous, a substance must be sufficiently volatile for its molecules to be given off and carried into the nostrils by air currents. The solubility of the substance also seems to play a role; chemicals that are soluble in water or fat tend to be strong odorants, although many of them are inodorous. No unique chemical or physical property that can be said to elicit the experience of odour has yet been defined.

Only seven of the chemical elements are odorous: fluorine, chlorine, bromine, iodine, oxygen (as ozone), phosphorus, and arsenic. Most odorous substances are organic (carbon-containing) compounds in which both the arrangement of atoms within the molecule as well as the particular chemical groups that comprise the molecule influence odour. Stereoisomers (*i.e.*, different spatial arrangements of the same molecular components) may have different odours. On the other hand, a series of different molecules that derive from benzene all have a similar odour. It is of historic interest that the first benzene derivatives studied by chemists were found in pleasant-smelling substances from plants (such as oil of wintergreen or oil of anise), and so the entire class of these compounds was labelled aromatic. Subsequently, other so-called aromatic compounds were identified that have less attractive odours.

The scent of flowers and roots (such as ginger) depends upon the presence of minute quantities of highly odorous essential oils. Although the major odour constitutents can be identified by chemical analysis, some botanical essences are so complex that their odours can be duplicated only by adding them in small amounts to synthetic formulations.

Odour sensitivity. In spite of the relative inaccessibility of the human olfactory receptor cells, odour stimuli can be detected at extremely low concentrations. Olfaction is said

to be 10,000 times more sensitive than taste. A human threshold value for such a well-known odorant as ethyl mercaptan (found in rotten meat) has been cited in the range of 1/400,000,000th of a milligram per litre of air. A just-noticeable difference in odour intensity may be apparent when there is a 20 percent increase in odorant strength, but at low concentrations as much as 100 percent increase in concentration may be required. Temperature influences the strength of an odour by affecting the volatility and hence the emission of odorous particles from the source; humidity also affects odour for the same reasons. Hunting dogs can follow a spoor (odour trail) most easily when high humidity retards evaporation and dissipation of the odour. Perfumes contain chemicals called fixatives, added to retard evaporation of the more volatile constituents. The temporary anosmia (absence of sense of smell) following colds in the nose may be complete or partial; in the latter case, only the odours of certain substances are affected. Paranosmia (change in perceived odour quality) also may occur during respiratory infections. Changes in sensitivity are reported to occur in women during the menstrual cycle, particularly in regard to certain odorants (steroids) related to sex hormones. Olfactory sensitivity also is said to become more acute during hunger.

Factors affecting odour sensitivity

Human adaptation to odours is so striking that the stench of a slaughterhouse or chemical laboratory ceases to be a nuisance after a few minutes have passed. Olfactory adaptation, as measured by a rise in threshold, is especially pronounced for stronger odours. Cross adaptation (between different odours) may take place; thus, eucalyptus oil may be difficult to detect after one becomes adapted to the smell of camphor.

Adaptation long was regarded solely as the result of changes in the olfactory receptor; however, electrical recordings show that the receptor cells in the nose seem to adapt only partially. Rhythmic discharges continue in the olfactory bulb long after the experimenter ceases to detect the odour that is stimulating the experimental animal. Apparently, some olfactory adaptation may occur in the brain as well as in the sense organ.

Odour blending and flavour. The ancient art of perfumery and the modern science of odour control depend on mixing and blending. Two different odours presented together may be readily identifiable. The more they resemble each other, the greater will be the tendency to blend; yet trained workers usually can discriminate the components of a successfully blended perfume. The substantially greater intensity of one odour may mask another. Masking, however, is less effective than chemical conversion or physical collection (removal) as a basis for odour control. Some odorants also may be removed by passing air through activated charcoal.

Flavours and the sense of smell

The distinctive flavours of food are known largely through the sense of smell. Flavour is the composite of experiences from many senses, but the aroma of roast beef and the delicate bouquet of wine are mainly olfactory in origin. Flavour technology has assumed a well-entrenched place in the food industry; it is common practice for manufacturers to use flavour panels of several trained members to judge the flavour of new food products. Using psychophysical methods with adequate statistical control, such panels are very reliable and can detect qualities so subtle that they defy the methods of physics and chemistry.

In using large, untrained consumer panels, the emphasis is on acceptability, often an emotional reaction to the product. Such untrained judgments are unstable and may vary among individuals because of idiosyncrasy or differences in experience. For example, the strong cheese odour so palatable to the gourmet can produce revulsion in the uninitiated.

Effects on behaviour. Recognition of friend or foe by social insects may depend upon olfactory cues: certain ants attack their own kind furiously if deprived of the sense of smell by amputation of their antennae; bees entering a strange hive are put to death because the scent of a foreign hive clings to them. Bees also have a specialized organ on the end of the abdomen that deposits a scent on a newly discovered food source to guide other foraging workers. The scents of flowers attest to the evolutionary importance of odour; those flowers that attract insects most efficiently are the likeliest to be pollinated and to reproduce their kind.

The effect of odours on the sexual behaviour of invertebrates is most striking; a female moth, for example, was observed to attract more than 100 males during relatively brief observation periods of about six hours. The physiological basis for the attraction can be traced to specific odour attractants (pheromones), molecules produced by the scent glands of the female. Pheromones also function in defense and as alarms of impending danger. The specific sex attractant of the female silk moth (*Bombyx*) has been identified and synthesized. It has been found that the antennae of the male silk moth contain olfactory receptors specifically sensitive to the female pheromone but that the females have no receptors to detect their own attractiveness.

Mammals in the wild state appear to utilize their odour glands for sexual attraction. Laboratory rats show a preference for the branch of a maze that has been scented with the odour of a sexually receptive female. It is likely that some rudiments of these effects operate in man. The most sexually provocative perfumes have a high proportion of musk or musklike odour. Genuine musk is derived from the sexual glands of the musk deer and is chemically related to human sex hormones; odour sensitivity in humans varies with the menstrual cycle.

Among laboratory animals the secretion of reproductive hormones can be markedly influenced by odour stimulation. This seems to be an innate physiological process rather than the result of learning. A most dramatic effect (pregnancy block) is observed when the odour of a strange male is presented to a recently mated female. The normal hormonal changes following copulation are blocked under these conditions, and the fertilized egg fails to survive. A related study of the periodicity and length of menstrual cycle in women exposed to the normal odours of men suggests there may be similar effects among people. Human behaviour, molded and shaped by custom and culture though it is, has many of its roots in his basic sensual appetites. (C.Pf.)

HUMAN VISION: STRUCTURE AND FUNCTION OF THE EYE

Anatomy of the visual apparatus

STRUCTURES AUXILIARY TO THE EYE

The orbit. The eye is protected from mechanical injury by being enclosed in a socket, or orbit, which is made up of portions of several of the bones of the skull to form a four-sided pyramid the apex of which points back into the head. Thus, the floor of the orbit is made up of parts of the maxilla, zygomatic, and palatine bones, while the roof is made up of the orbital plate of the frontal bone and, behind this, by the lesser wing of the sphenoid. The optic foramen, the opening through which the optic nerve runs back into the brain and the large ophthalmic artery enters the orbit, is at the nasal side of the apex; the superior orbital fissure is a larger hole through which pass large veins and nerves. These nerves may carry nonvisual sensory messages—*e.g.*, pain—or they may be motor nerves controlling the muscles of the eye. There are other fissures and canals transmitting nerves and blood vessels. The eyeball and its functional muscles are surrounded by a layer of orbital fat that acts much like a cushion, permitting a smooth rotation of the eyeball about a virtually fixed point, the centre of rotation. The protrusion of the eyeballs—proptosis—in exophthalmic goitre is caused by the collection of fluid in the orbital fatty tissue.

The eyelids. It is vitally important that the front surface of the eyeball, the cornea, remain moist. This is achieved by the eyelids, which during waking hours sweep

the secretions of the lacrimal apparatus and other glands over the surface at regular intervals and which during sleep cover the eyes and prevent evaporation. The lids have the additional function of preventing injuries from foreign bodies, through the operation of the blink reflex. The lids are essentially folds of flesh covering the front of the orbit and, when the eye is open, leaving an almond-shaped aperture. The points of the almond are called canthi; that nearest the nose is the inner canthus, and the other is the outer canthus (Figure 36). The lid may be

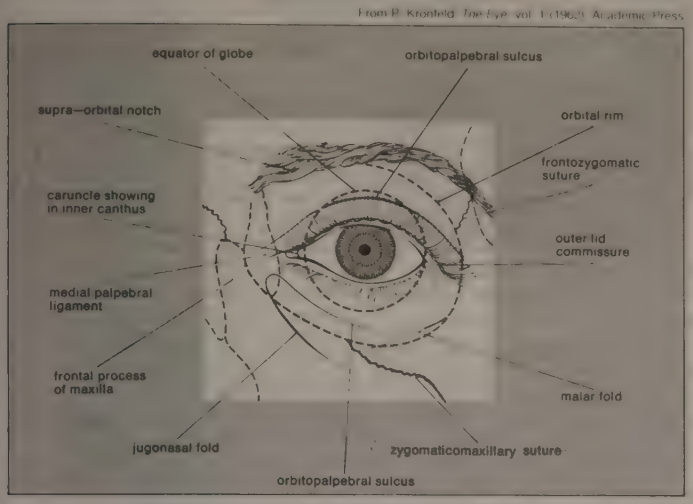


Figure 36: Frontal view of the eye and its related structures (see text).

divided into four layers: (1) the skin, containing glands that open onto the surface of the lid margin, and the eyelashes; (2) a muscular layer containing principally the orbicularis oculi muscle, responsible for lid closure; (3) a fibrous layer that gives the lid its mechanical stability, its principal portions being the tarsal plates, one in each lid, which border directly upon the opening between the lids, called the palpebral aperture; and (4) the innermost layer of the lid, a portion of the conjunctiva. The conjunctiva is a mucous membrane that serves to attach the eyeball to the orbit and lids but permits a considerable degree of rotation of the eyeball in the orbit.

The layers in the lids

The conjunctiva. The conjunctiva lines the lids and then bends back over the surface of the eyeball, constituting an outer covering to the forward part of this and terminating at the transparent region of the eye, the cornea. The portion that lines the lids is called the palpebral portion of the conjunctiva; the portion covering the white of the eyeball is called the bulbar conjunctiva. Between the bulbar and the palpebral conjunctiva there are two loose, redundant portions forming recesses that project back toward the equator of the globe. These recesses are called the upper and lower fornices, or conjunctival sacs; it is the looseness of the conjunctiva at these points that makes movements of lids and eyeball possible. (The ophthalmologist finds the lower conjunctival sac a useful cavity in which to place drops containing drugs. He accomplishes this by merely pulling the outer lid away from the globe. The drops are retained in the cavity long enough to act directly on the cornea and to diffuse through this into the internal structures of the eye.)

The fibrous layer. The fibrous layer, which gives the lid its mechanical stability, is made up of the thick, and relatively rigid, tarsal plates, bordering directly on the palpebral aperture, and the much thinner palpebral fascia, or sheet of connective tissue; the two together are called the septum orbitale. When the lids are closed, the whole opening of the orbit is covered by this septum. Two ligaments, the medial and lateral palpebral ligaments, attached to the orbit and to the septum orbitale, stabilize the position of the lids in relation to the globe. The medial ligament, by far the stronger, is well illustrated in Figure 36.

The muscles. Closure of the lids is achieved by contraction of the orbicularis muscle, a single oval sheet of muscle extending from the regions of the forehead and face and surrounding the orbit into the lids. It is divided into orbital and palpebral portions, and it is essentially the palpebral portion, within the lid, that causes lid closure. The palpebral portion passes across the lids from a ligament called the medial palpebral ligament and from the neighbouring bone of the orbit in a series of half ellipses that meet outside the outer corner of the eye, the lateral canthus, to form a band of fibres called the lateral palpebral raphe. Additional parts of the orbicularis have been given separate names—namely, Horner's muscle and the muscle of Riolan; they come into close relation with the lacrimal apparatus and assist in drainage of the tears. The muscle of Riolan, lying close to the lid margins, doubtless contributes to keeping the lids in close apposition, an important feature for maintaining the junction watertight. The orbital portion of the orbicularis is not normally concerned with blinking, which may be carried out entirely by the palpebral portion; however, it is concerned with closing the eyes tightly. The skin of the forehead, temple, and cheek is then drawn toward the medial (nose) side of the orbit, and the radiating furrows, formed by this action of the orbital portion, eventually lead to the so-called crow's feet of elderly persons. It must be appreciated that the two portions can be activated independently; thus, the orbital portion may contract, causing a furrowing of the brows that reduces the amount of light entering from above, while the palpebral portion remains relaxed and allows the eyes to remain open.

Horner's muscle and the muscle of Riolan

Opening of the eye is not just the result of passive relaxation of the orbicularis muscle but also is the effect of the contraction of the levator palpebrae superioris muscle of the upper lid. This muscle takes origin with the extraocular muscles at the apex of the orbit (the back of the eye socket) as a narrow tendon and runs forward into the upper lid as a broad tendon, the levator aponeurosis, which is attached to the forward surface of the tarsus and the skin covering the upper lid. Contraction of the muscle causes elevation of the upper eyelid. The nervous connections of this muscle are closely related to those of the extraocular muscle required to elevate the eye, so that when the eye looks upward the upper eyelid tends to move up in unison.

The orbicularis and levator are striped muscles under voluntary control. The lids contain, in addition, unstriped (involuntary) muscle fibres that are activated by the sympathetic division of the autonomic system and tend to widen the palpebral fissure (the eye opening) by elevation of the upper, and depression of the lower, lid.

In addition to the muscles already described, other facial muscles often cooperate in the act of lid closure or opening. Thus, the corrugator supercilii muscles pull the eyebrows toward the bridge of the nose, making a projecting "roof" over the medial angle of the eye and producing characteristic furrows in the forehead; the roof is used primarily to protect the eye from the glare of the sun. The pyramidalis, or procerus, muscles occupy the bridge of the nose; they arise from the lower portion of the nasal bones and are attached to the skin of the lower part of the forehead on either side of the midline; they pull the skin into transverse furrows. In lid opening, the frontalis

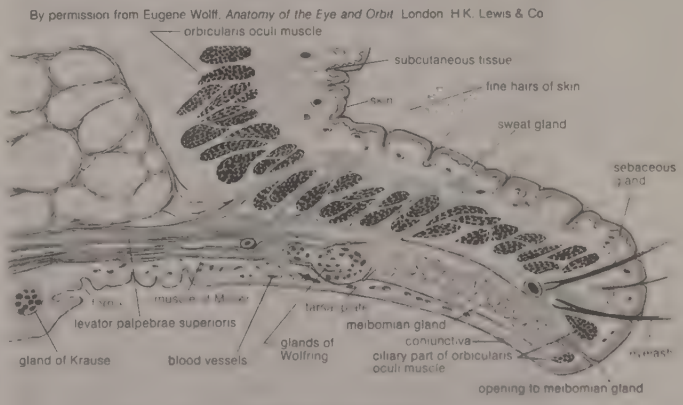


Figure 37: Vertical section through upper lid (see text).

muscle, arising high on the forehead, midway between the coronal suture, a seam across the top of the skull, and the orbital margin, is attached to the skin of the eyebrows. Contraction therefore causes the eyebrows to rise and opposes the action of the orbital portion of the orbicularis; the muscle is especially used when one gazes upward. It is also brought into action when vision is rendered difficult either by distance or the absence of sufficient light.

The skin. The outermost layer of the lid is the skin, with features not greatly different from skin on the rest of the body, with the possible exception of large pigment cells, which, although found elsewhere, are much more numerous in the skin of the lids. The cells may wander, and it is these movements of the pigment cells that determine the changes in coloration seen in some people with alterations in health. The skin has sweat glands and hairs. As the junction between skin and conjunctiva is approached, the hairs change their character to become eyelashes (Figure 37).

Tear glands

The glandular apparatus. The eye is kept moist by secretions of the lacrimal glands (tear glands). These almond-shaped glands under the upper lids extend inward from the outer corner of each eye. Each gland has two portions. One portion is in a shallow depression in the part of the eye socket formed by the frontal bone. The other portion projects into the back part of the upper lid. The ducts from each gland, three to 12 in number, open into the superior conjunctival fornix, or sac. From the fornix, the tears flow down across the eye and into the puncta lacrimalia, small openings at the margin of each eyelid near its inner corner. The puncta are openings into the lacrimal ducts; these carry the tears into the lacrimal sacs, the dilated upper ends of the nasolacrimal ducts, which carry the tears into the nose.

The evaporation of the tears as they flow across the eye is largely prevented by the secretion of oily and mucous material by other glands. Thus, the meibomian, or tarsal glands, consist of a row of elongated glands extending through the tarsal plates; they secrete an oil that emerges onto the surface of the lid margin and acts as a barrier for the tear fluid, which accumulates in the grooves between the eyeball and the lid barriers.

Extraocular muscles. Six muscles outside the eye govern its movements. These muscles are the four rectus muscles—the inferior, medial, lateral, and superior recti—and the superior and inferior oblique muscles. The rectus muscles arise from a fibrous ring that encircles the optic nerve at the optic foramen, the opening through which the nerve passes, and are attached to the sclera, the opaque portion of the eyeball, in front of the equator, or widest part, of the eye. The superior oblique muscle arises near the rim of the optic foramen and somewhat nearer the nose than the origin of the rectus medialis. It ends in a rounded tendon that passes through a fibrous ring, the trochlea, that is attached to the frontal bone. The trochlea acts as a pulley. The tendon is attached to the sclera back of the equator of the eye (Figure 38).

The inferior oblique muscle originates from the floor of the orbit, passes under the eyeball like a sling, and is attached to the sclera between the attachments of the superior and lateral rectus muscles. The rectus muscles direct the gaze upward and downward and from side to side. The inferior oblique muscle tends to direct the eye upward, and the superior oblique to depress the eye; because of the obliqueness of the pull, each causes the eye to roll, and in an opposite direction.

The oblique muscles are strictly antagonistic to each other, but they work with the vertical rectus muscles in so far as the superior rectus and inferior oblique both tend to elevate the gaze and the inferior rectus and superior oblique both tend to depress the gaze. The superior and inferior recti do not produce a pure action of elevation or depression because their plane of action is not exactly vertical; in consequence, as with the obliques, they cause some degree of rolling, but by no means so great as that caused by the obliques; the direction of rolling caused by the rectus muscle is opposite to that of its synergistic oblique; the superior rectus causes the eye to roll inward, and the inferior oblique outward.

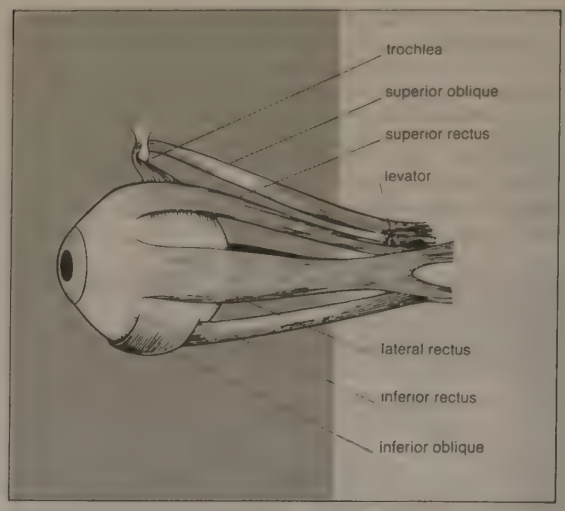


Figure 38: The extraocular muscles.
From P. Kronfeld, *The Eye*, vol. 1 (1962) Academic Press

THE EYE

General description. The eyeball is not a simple sphere but can be viewed as the result of fusing a small portion of a small, strongly curved sphere with a large portion of a large, not so strongly curved sphere (Figure 39). The small piece, occupying about one-sixth of the whole, has a radius of eight millimetres (0.3 inch); it is transparent and is called the cornea; the remainder, the scleral segment, is opaque and has a radius of 12 millimetres (0.5 inch). The

Visible parts of the eye

From H. Davson, M.D., *Physiology of the Eye*

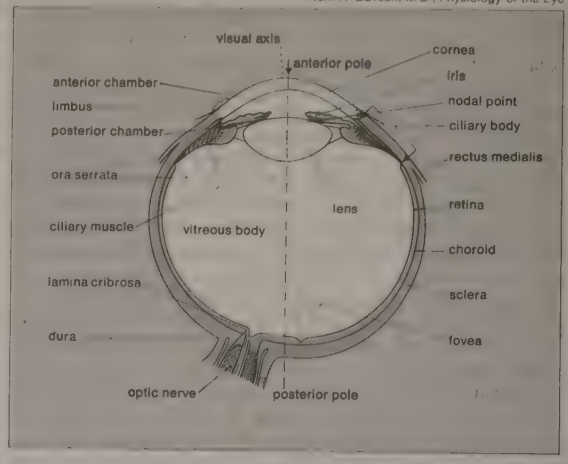


Figure 39: Horizontal section of the eye.

ring where the two areas join is called the limbus. Thus, on looking directly into the eye from in front one sees the white sclera surrounding the cornea; because the latter is transparent one sees, instead of the cornea, a ring of tissue lying within the eye, the iris. The iris is the structure that determines the colour of the eye. The centre of this ring is called the pupil. It appears dark because the light passing into the eye is not reflected back to any great extent. By use of an ophthalmoscope, an instrument that permits the observer to illuminate the interior of the eyeball while observing through the pupil, the appearance of the interior lining of the globe can be made out; this is called the fundus; it is characterized by the large blood vessels that suppy blood to the retina; these are especially distinct as they cross over the pallid optic disk, or papilla, the region where the optic nerve fibres leave the globe.

The dimensions of the eye are reasonably constant, varying among individuals by only a millimetre or two; the sagittal (vertical) diameter is about 24 millimetres (about one inch) and is usually less than the transverse diameter. At birth the sagittal diameter is about 16 to 17 millimetres (about 0.65 inch); it increases rapidly to about 22.5 to 23 millimetres (about 0.89 inch) by the age of three years; between three and 13 the globe attains its full size. The

weight is about 7.5 grams (.25 ounce), and its volume 6.5 millilitres (0.4 cubic inch).

The eye is made up of three coats, which enclose the optically clear aqueous humour, lens, and vitreous body (Figure 39). The outermost coat consists of the cornea and the sclera; the middle coat contains the main blood supply to the eye and consists, from the back forward, of the choroid, the ciliary body, and the iris. The innermost layer is the retina, lying on the choroid and receiving most of its nourishment from the vessels within the choroid, the remainder of its nourishment being derived from the retinal vessels that lie on its surface and are visible in the ophthalmoscope. The ciliary body and iris have a very thin covering, the ciliary epithelium and posterior epithelium of the iris, which is continuous with the retina.

Within the cavities formed by this triple-layered coat there are the crystalline lens, suspended by fine transparent fibres—the suspensory ligament or zonule of Zinn—from the ciliary body; the aqueous humour, a clear fluid filling the spaces between the cornea and the lens and iris; and the vitreous body, a clear jelly filling the much larger cavity enclosed by the sclera, the ciliary body, and the lens. The anterior chamber of the eye is defined as the space between the cornea and the forward surfaces of the iris and lens, while the posterior chamber is the much smaller space between the rear surface of the iris and the ciliary body, zonule, and lens; the two chambers both contain aqueous humour and are in connection through the pupil.

Outer and middle tunics of the globe. *The outermost coat.* The outermost coat is made up of the cornea and the sclera. The cornea is the transparent window of the eye. It contains five distinguishable layers; the epithelium, or outer covering; Bowman's membrane; the stroma, or supporting structure; Descemet's membrane; and the endothelium, or inner lining. Up to 90 percent of the thickness of the cornea is made up of the stroma. The epithelium, which is a continuation of the epithelium of the conjunctiva, is itself made up of about six layers of cells. The superficial layer is continuously being shed, and the layers are renewed by multiplication of the cells in the innermost, or basal, layer.

The layers of the cornea

The stroma appears as a set of lamellae, or plates, running parallel with the surface and superimposed on each other like the leaves of a book; between the lamellae lie the corneal corpuscles, cells that synthesize new collagen (connective tissue protein) essential for the repair and maintenance of this layer. The lamellae are made up of microscopically visible fibres that run parallel to form sheets; in successive lamellae the fibres make a large angle with each other. The lamellae in man are about 1.5 to 2.5 microns (one micron = 0.001 millimetre) thick, so that there are about 200 lamellae in the human cornea. The fibrous basis of the stroma is collagen.

Immediately above the stroma, adjacent to the epithelium, is Bowman's membrane, about eight to 14 microns thick; in the electron microscope it is evident that it is really stroma, but with the collagen fibrils not arranged in the orderly fashion seen in the rest of the stroma.

Beneath the stroma are Descemet's membrane and the endothelium. The former is about five to 10 microns thick and is made up of a different type of collagen from that in the stroma; it is secreted by the cells of the endothelium, which is a single layer of flattened cells. There is apparently no continuous renewal of these cells as with the epithelium, so that damage to this layer is a more serious matter.

The sclera is essentially the continuation backward of the cornea, the collagen fibres of the cornea being, in effect, continuous with those of the sclera. The sclera is pierced by numerous nerves and blood vessels; the largest of these holes is that formed by the optic nerve, the posterior scleral foramen. The outer two-thirds of the sclera in this region continue backward along the nerve to blend with its covering, or dural sheath—in fact, the sclera may be regarded as a continuation of the dura mater, the outer covering of the brain. The inner third of the sclera, combined with some choroidal tissue, stretches across the opening, and the sheet thus formed is perforated to permit the passage of fasciculi (bundles of fibres) of the optic nerve. This region is called the lamina cribrosa (Figure 39). The blood vessels of the sclera are largely confined to a superficial layer of tissue, and these, along with the conjunctival vessels, are responsible for the bright redness of the inflamed eye. As with the cornea, the innermost layer is a single layer of endothelial cells; above this is the lamina fusca, characterized by large numbers of pigment cells.

The most obvious difference between the opaque sclera and the transparent cornea is the irregularity in the sizes and arrangement of the collagen fibrils in the sclera by contrast with the almost uniform thickness and strictly parallel array in the cornea; in addition, the cornea has a much higher percentage of mucopolysaccharide (a carbohydrate that has among its repeating units a nitrogenous sugar, hexosamine) as embedding material for the collagen fibrils. It has been shown that the regular arrangement of the fibrils is, in fact, the essential factor leading to the transparency of the cornea.

Differences between sclera and cornea

When the cornea is damaged—*e.g.*, by a virus infection—the collagen laid down in the repair processes is not regularly arranged, with the result that an opaque patch called a leukoma, may occur.

When an eye is removed, or a man dies, the cornea soon loses its transparency, becoming hazy; this is due to the taking in of fluid from the aqueous humour, the cornea becoming thicker as it becomes hazier. The cornea can be made to reassume its transparency by maintaining it in a warm, well-aerated chamber, at about 31° C (88° F, its normal temperature); associated with this return of transparency is a loss of fluid.

Modern studies have shown that, under normal conditions, the cornea tends to take in fluid, mainly from the aqueous humour and from the small blood vessels at the limbus, but this is counteracted by a pump that expels the fluid as fast as it enters. This pumping action depends on an adequate supply of energy, and any situation that prejudices this supply causes the cornea to swell—the pump fails, or works so slowly that it cannot keep pace with the leak. Death is one cause of the failure of the pump, but this is primarily because of the loss of temperature; place the dead eye in a warm chamber and the reserves of metabolic energy it contains in the form of sugar and glycogen are adequate to keep the cornea transparent for 24 hours or more. When it is required to store corneas for grafting, as in an eye bank, it is best to remove the cornea from the globe to prevent it from absorbing fluid from the aqueous humour. The structure responsible for the pumping action is almost certainly the endothelium, so that damage to this lining can lead to a loss of transparency with swelling.

The cornea is exquisitely sensitive to pain. This is mediated by sensory nerve fibres, called ciliary nerves, that run just underneath the endothelium; they belong to the ophthalmic branch of the fifth cranial nerve, the large sensory nerve of the head. The ciliary nerves leave the globe through holes in the sclera, not in company with the optic nerve, which is concerned exclusively with responses of the retina to light.

The uvea. The middle coat of the eye is called the uvea (from the Latin for "grape") because the eye looks like a reddish-blue grape when the outer coat has been dissected away. The posterior part of the uvea, the choroid, is essentially a layer of blood vessels and connective tissue sandwiched between the sclera and the retina. The forward portion of the uvea, the ciliary body and iris, is more complex, containing as it does the ciliary muscle and the sphincter and dilator of the pupil.

Ciliary body and iris

The blood supply to the human eye is twofold, consisting of the retinal and uveal circulations, both of which derive from branches of the ophthalmic artery. The two systems of blood vessels differ in that the retinal vessels, which supply nutrition to the innermost layers of the retina, derive from a branch of the ophthalmic artery, called the central artery of the retina, that enters the eye with the optic nerve, while the uveal circulation, which supplies the middle and outer layers of the retina as well as the uvea, is derived from branches of the ophthalmic artery that penetrate the globe independently of the optic nerve.

The ciliary body is the forward continuation of the

choroid. It is a muscular ring, triangular in horizontal section, beginning at the region called the ora serrata and ending, in front, as the root of the iris (Figure 40). The surface is thrown into folds, called ciliary processes, the whole being covered by the ciliary epithelium, which is a double layer of cells; the layer next to the vitreous body (see below), called the inner layer, is transparent, while the outer layer, which is continuous with the pigment

From P. Kronfeld, *The Eye*, vol. 1 (1962), Academic Press

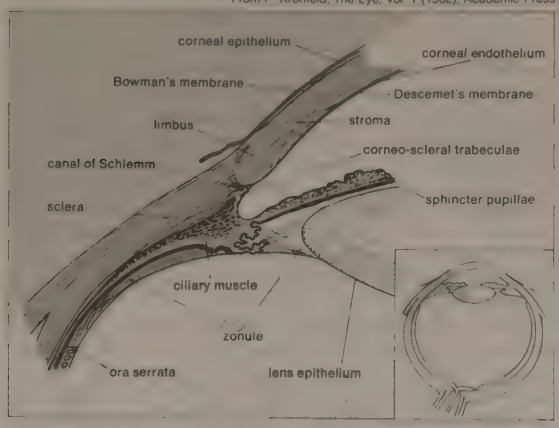


Figure 40: The anteronasal portion of a horizontal (meridional) section through a right eye. Shaded area of inset locates magnified portion. (See text.)

epithelium of the retina, is heavily pigmented. These two layers are to be regarded embryologically as the forward continuation of the retina, which terminates at the ora serrata. Their function is to secrete the aqueous humour.

The ciliary muscle is an unstriped, involuntary, muscle concerned with alterations in the adjustments of focus—accommodation—of the optical system; the fibres run both across the muscle ring and circularly, and the effect of their contraction is to cause the whole body to move forward and to become fatter, so that the suspensory ligament that holds the lens in place is loosened.

The most forward portion of the uvea is the iris. This is the only portion that is visible to superficial inspection, appearing as a perforated disc, the central perforation, or pupil, varying in size according to the surrounding illumination and other factors. A prominent feature is the collarette at the inner edge, representing the place of attachment of the embryonic pupillary membrane that, in embryonic life, covers the pupil. As with the ciliary body, with which it is anatomically continuous, the iris consists of several layers: namely, an anterior layer of endothelium, the stroma; and the posterior iris epithelium. The stroma contains the blood vessels and the sphincter and dilator muscles; in addition, the stroma contains pigment cells that determine the colour of the eye. Posteriorly, the stroma is covered by a double layer of epithelium, the continuation forward of the ciliary epithelium; here, however, both layers are heavily pigmented and serve to prevent light from passing through the iris tissue, confining the optical pathway to the pupil. The pink iris of the albino is the result of the absence of pigment in these layers. The cells of the anterior layer of the iris epithelium have projections that become the fibres of the dilator muscle; these projections run radially, so that when they contract they pull the iris into folds and widen the pupil; by contrast, the fibres of the sphincter pupillae muscle run in a circle around the pupil, so that when they contract the pupil becomes smaller.

Usually, a baby belonging to the white races is born with blue eyes because of the absence of pigment cells in the stroma; the light reflected back from the posterior epithelium, which is blue because of scattering and selective absorption, passes through the stroma to the eye of the observer. As time goes on, pigment is deposited, and the colour changes; if much pigment is laid down the eye becomes brown or black, if little, it remains blue or gray.

The inner tunic of the globe. The inner tunic of the rear portion of the globe, as far forward as the ciliary body, is the retina, including its epithelia or coverings. These epithelia continue forward to line the remainder of the globe.

The epithelia. Separating the choroid (the middle tunic of the globe) from the retina proper is a layer of pigmented cells, the pigment epithelium of the retina; this acts as a restraining barrier to the indiscriminate diffusion of material from the blood in the choroid to the retina. The retina ends at the ora serrata, where the ciliary body begins (Figure 39). The pigment epithelium continues forward as a pigmented layer of cells covering the ciliary body; farther forward still, the epithelium covers the posterior surface of the iris and provides the cells that constitute the dilator muscle of this diaphragm. Next to the pigment epithelium of the retina is the neuroepithelium, or rods and cones (see below). Their continuation forward is represented by a second layer of epithelial cells covering the ciliary body, so that by the ciliary epithelium is meant the two layers of cells that are the embryological equivalent of the retinal pigment epithelium and the receptor layer (rods and cones) of the retina. This unpigmented layer of the ciliary epithelium is continued forward over the back of the iris, where it acquires pigment and is called the posterior iris epithelium.

The pigment epithelium

The retina. The retina is the part of the eye that receives the light and converts it into chemical energy. The chemical energy activates nerves that conduct the messages out of the eye into the higher regions of the brain. The retina is a complex nervous structure, being, in essence, an outgrowth of the forebrain.

Ten layers of cells in the retina can be seen microscopically. In general, there are four main layers: (1) Next to the choroid is the pigment epithelium, already mentioned. (2) Beneath the epithelium is the layer of rods and cones, the light-sensitive cells. The changes induced in the rods and cones by light are transmitted to (3) a layer of neurons (nerve cells) called the bipolar cells, which are analogous to the sensory neurons that carry messages from the touch and heat receptors of the skin and transmit them to the cells of the spinal cord or the medulla (the part of the brain that is a continuation of the spinal cord). These bipolar cells connect with (4) the innermost layer of neurons, the ganglion cells; and the transmitted messages are carried out of the eye along their projections, or axons, which constitute the optic nerve fibres. Thus, the optic nerve is really a central tract, rather than a nerve, connecting two regions of the nervous system, namely, the layer of bipolar cells, and the cells of the lateral geniculate body, the latter being a visual relay station in the diencephalon (the rear portion of the forebrain).

The arrangement of the retinal cells in an orderly manner gives rise to the outer nuclear layer (layer 4 in Figure 41), containing the nuclei of the rods and cones; the inner nuclear layer (layer 6), containing the nuclei and perikarya (main cell bodies outside the nucleus) of the bipolar cells, and the ganglion cell layer (layer 8), containing the corresponding structures of the ganglion cells. The plexiform layers are regions in which the neurons make their interconnections. Thus, the outer plexiform layer (layer 5) contains the rod and cone projections terminating as the rod spherule and cone pedicle; these make connections with the dendritic processes of the bipolar cells, so that changes produced by light in the rods and cones are transmitted by way of these connections to the bipolar cells. (The dendritic process of a nerve cell is the projection that receives nerve impulses to the cell; the axon is the projection that carries impulses from the cell.) In the inner plexiform layer (layer 7) are the axons of the bipolar cells and the dendritic processes of the ganglion and amacrine cells (see below). The association is such as to allow messages in the bipolar cells to be transmitted to the ganglion cells, the messages then passing out along the axons of the ganglion cells as optic nerve messages.

The photosensitive cells are, in the human and in most vertebrate retinas, of two kinds, called rods and cones, the rods being usually much thinner than the cones but both being built up on the same plan. The light-sensitive pigment is contained in the outer segment (layer 2), which rests on the pigment epithelium (layer 1). Through the other end, called the synaptic body, effects of light are

The rods and cones

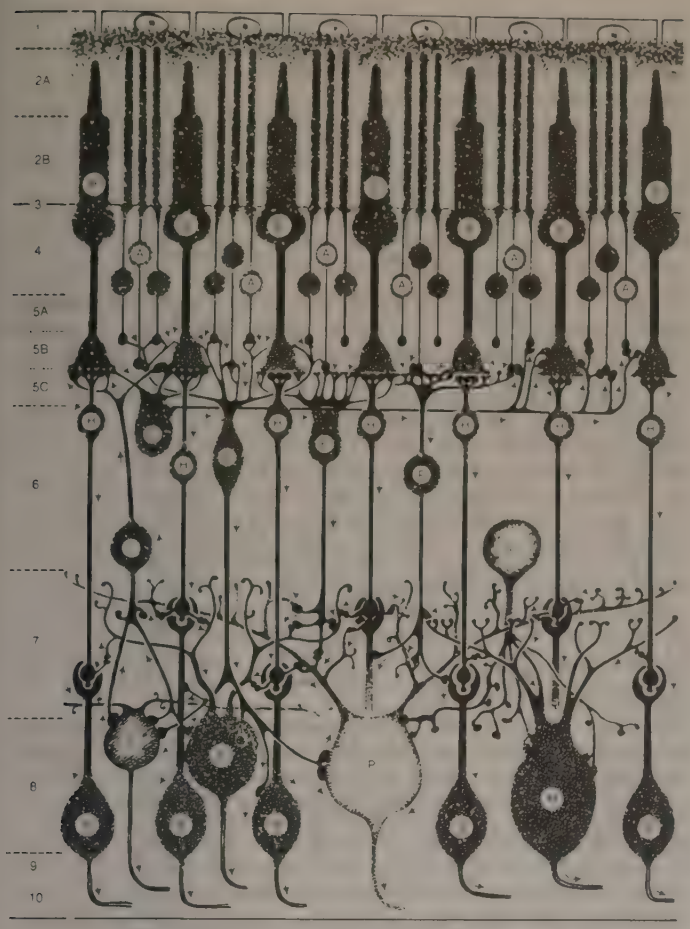


Figure 41: *The retinal cells directly involved in the visual process.* (A) Rod and (B) cone cells, the photoreceptors. (C) Horizontal cells. (D) Mop, (E) brush, (F) flat-topped, (H) midget, and (I) centrifugal varieties of bipolar cells. (L) Several varieties of amacrine cells. (M) Parasol, (N) shrub, (O) small diffuse, (P) garland, and (S) midget ganglion varieties of ganglion cells.

By courtesy of Stephen L. Polyak

transmitted to the bipolar and horizontal cells. When examined in the electron microscope, the outer segments of the rods and cones are seen to be composed of stacks of disks, apparently made by the infolding of the limiting membrane surrounding the outer segment; the visual pigment, located on the surfaces of these disks, is thus spread over a very wide area, and this contributes to the efficiency with which light is absorbed by the visual cell.

The arrangement of the retina makes it necessary for light to pass through the layers not sensitive to light first before it reaches the light-sensitive rods and cones. The optical disadvantages of this arrangement are largely overcome by the development of the fovea centralis, a localized region of the retina, close to the optic axis of the eye, where the inner layers of the retina are absent. The result is a depression, the foveal pit, where light has an almost unrestricted passage to the light-sensitive cells. It is essentially this region of the retina that is employed for accurate vision, the eyes being directed toward the objects of regard so that their images fall in this restricted region. If the object of interest is large, so as to subtend a large angle, then the eye must move rapidly from region to region so as to bring their images successively onto the fovea; this is typically seen during reading. In the central region of the fovea there are cones exclusively; toward its edges, rods also occur, and as successive zones are reached the proportion of rods increases while the absolute density of packing of the receptors tends to decrease. Thus, the central fovea is characterized by an exclusive population of very densely packed cones; here, also, the cones are very thin and in form very similar to rods. The region surrounding the fovea is called the parafovea; it stretches about 1,250 microns from the centre of the fovea, and it is here that the highest density of rods occurs. Surrounding the parafovea, in turn, is the perifovea, its outermost edge being 2,750 microns from the centre of the fovea; here the density of cones is still further diminished, the number being only 12 per hundred microns compared with 50 per hundred microns in the most central region of the fovea. In the whole human retina there are said to be about 7,000,000 cones and from 75,000,000 to 150,000,000 rods.

Retinal organization

The fovea is sometimes referred to as the macula lutea ("yellow spot"); actually this term defines a rather vague area, characterized by the presence of a yellow pigment in the nervous layers, stretching over the whole central retina; *i.e.*, the fovea, parafovea, and perifovea.

The blind spot in the retina corresponds to the optic papilla, the region on the nasal side of the retina through which the optic nerve fibres pass out of the eye.

Although the rods and cones may be said to form a mosaic, the retina is not organized in a simple mosaic fashion in the sense that each rod or cone is connected to a single bipolar cell that itself is connected to a single ganglion cell. There are only about 1,000,000 optic nerve fibres, while there are at least 150,000,000 receptors, so that there must be considerable convergence of receptors on the optic pathway. This means that there will be considerable mixing of messages. Furthermore, the retina contains additional nerve cells besides the bipolar and ganglion cells; these, the horizontal and amacrine cells, operate in the horizontal direction, allowing one area of the retina to influence the activity of another. In this way, for example, the messages from one part of the retina may be suppressed by a visual stimulus falling on another, an important element in the total of messages sent to the higher regions of the brain. Finally, it has been argued that some messages may be running the opposite way; they are called centrifugal and would allow one layer of the retina to affect another, or higher regions of the brain to control the responses of the retinal neurons. In primates the existence of these centrifugal fibres has been finally disproved, but in such lower vertebrates as the pigeon, their existence is quite certain.

The pathway of the retinal messages through the brain is described later in this article; it is sufficient to state here that most of the optic nerve fibres in primates carry their messages to the lateral geniculate body, a relay station specifically concerned with vision. Some of the fibres separate from the main stream and run to a midbrain centre called the pretectal nucleus, which is a relay centre for pupillary responses to light.

Secretion and course of aqueous humour

The transparent media. Within the cavities enclosed by the three layers of the globe described above there are the aqueous humour in the anterior and posterior chambers; the crystalline lens behind the iris; and the vitreous body, which fills the large cavity behind the lens and iris (Figure 39).

The aqueous humour. The aqueous humour is a clear colourless fluid with a chemical composition rather similar to that of blood plasma (the blood exclusive of its cells) but lacking the high protein content of the latter. Its main function is to keep the globe reasonably firm. It is secreted continuously by the ciliary body into the posterior chamber, and flows as a gentle stream through the pupil into the anterior chamber, from which it is drained by way of a channel at the limbus; that is, the juncture of the cornea and the sclera. This channel, the canal of Schlemm, encircles the cornea and connects by small connector channels to the blood vessels buried in the sclera and forming the intrascleral plexus or network. From this plexus the blood, containing the aqueous humour, passes into more superficial vessels; it finally leaves the eye in the anterior ciliary veins. The relation of the canal of Schlemm to the aqueous humour is clear from Figure 40. The wall of the canal that faces the aqueous humour is very delicate and allows the fluid to percolate through by virtue of the relatively high pressure of the fluid within the eye. Obstruction of this exit, for example, if the iris is pushed forward to cover the wall of the canal, causes a sharp rise in the pressure within the eye, a condition that is known as glaucoma. Often the obstruction is not obvious, but is caused perhaps by a hardening of the tissue just adjacent to the wall of the canal—the trabecular

meshwork (Figure 40), in which case the rise of pressure is more gradual and insidious. Ultimately the abnormal pressure damages the retina and causes a variable degree of blindness. The normal intraocular pressure is about 15 millimetres (0.6 inch) of mercury above atmospheric pressure, so that if the anterior chamber is punctured by a hypodermic needle the aqueous humour flows out readily. Its function in maintaining the eye reasonably hard is seen by the collapse and wrinkling of the cornea when the fluid is allowed to escape. An additional function of the fluid is to provide nutrition for the crystalline lens and also for the cornea, both of which are devoid of blood vessels; the steady renewal and drainage serve to bring into the eye various nutrient substances, including glucose and amino acids, and to remove waste products of metabolism.

The vitreous body. The vitreous body is a jelly. It is remarkable for the small amount of solid matter required to give it this semisolid structure; the solid material is made up of a form of collagen, vitrosin, and a mucopolysaccharide, hyaluronic acid. Thus, its composition is rather similar to that of the cornea, but the proportion of water is much greater, about 98 percent or more, compared with about 75 percent for the cornea. The jelly is probably secreted by certain cells of the retina. In general, the vitreous body is devoid of cells, in contrast with the lens, which is packed tight with cells. Embedded in the surface of the vitreous body, however, there is a population of specialized cells, the hyalocytes of Balazs, which may contribute to the breakdown and renewal of the hyaluronic acid. The vitreous body serves to keep the underlying retina pressed against the choroid.

The crystalline lens. The lens is a transparent body, flatter on its anterior than on its posterior surface, and suspended within the eye by the zonular fibres of Zinn attached to its equator; its anterior surface is bathed by aqueous humour, and its posterior surface by the vitreous body. The lens is a mass of tightly packed transparent fibrous cells, the lens fibres, enclosed in an elastic collagenous capsule. The lens fibres are arranged in sheets that form successive layers; the fibres run from pole to pole of the lens, the middle of a given fibre being in the equatorial region. On meridional (horizontal) section, the fibres are cut longitudinally to give an onion-scale appearance, whereas a section at right-angles to this—an equatorial section—would cut all the fibres across, and the result would be to give a honeycomb appearance. The epithelium, covering the anterior surface of the lens under the capsule, serves as the origin of the lens fibres, both during embryonic and fetal development and during infant and adult life, the lens continuing to grow by the laying down of new fibres throughout life.

The visual process

THE WORK OF THE AUXILIARY STRUCTURES

The protective mechanisms. The first line of protection of the eyes is provided by the lids, which prevent access of foreign bodies and assist in the lubrication of the corneal surface. Lid closure and opening are accomplished by the orbicularis oculi and levator palpebri muscles; the orbicularis oculi operates on both lids, bringing their margins into close apposition in the act of lid closure. Opening results from relaxation of the orbicularis muscle and contraction of the levator palpebri of the upper lid; the smooth muscle of the upper lid, Müller's muscle, or the superior palpebral muscle, also assists in widening the lid aperture. The lower lid does not possess a muscle corresponding to the levator of the upper lid, and the only muscle available for causing an active lowering of the lid, required during the depression of the gaze, is the inferior palpebral muscle, which is analogous to the muscle of Müller of the upper lid (called the superior palpebral muscle). This inferior palpebral muscle is so directly fused with the sheaths of the ocular muscles that it provides cooperative action, opening of the lid on downward gaze being mediated, in effect, mainly by the inferior rectus.

Blinking

Innervation. The seventh cranial nerve—the facial nerve—supplies the motor fibres for the orbicularis muscle. The levator is innervated by the third cranial nerve—the oculomotor nerve—that also innervates some of the extraocular muscles concerned with rotation of the eyeball, including the superior rectus. The smooth muscle of the eyelids and orbit is activated by the sympathetic division of the autonomic system. The secretion of adrenaline during such states of excitement as fear would also presumably cause contraction of the smooth muscle, but it seems unlikely that this would lead to the protrusion of the eyes traditionally associated with extreme fear. It is possible that the widening of the lid aperture occurring in this excited state, and dilation of the pupil, create the illusion of eye protrusion.

Voluntary full closure of lids

Blinking is normally an involuntary act, but may be carried out voluntarily. The more vigorous "full closure" of the lids involves the orbital portion of the orbicularis muscle and may be accompanied by contraction of the facial muscles that have been described as accessory muscles of blinking: namely, the corrugator supercilii, which on contraction pulls the eyebrows toward the bridge of the nose; and the procerus or pyramidalis, which pulls the skin of the forehead into horizontal folds, acting as a protection when the eyes are exposed to bright light. The more vigorous full closure may be evoked as a reflex response.

Blink reflexes. Reflex blinking may be caused by practically any peripheral stimulus, but the two functionally significant reflexes are (1) that resulting from stimulation of the endings of the fifth cranial nerve in the cornea, lid, or conjunctiva—the sensory blink reflex, or corneal reflex—and (2) that caused by bright light—the optical blink reflex. The corneal reflex is rapid (0.1 second reflex time) and is the last to disappear in deepening anesthesia, impulses being relayed from the nucleus of the fifth nerve to the seventh cranial nerve, which transmits the motor impulses. The reflex is said to be under the control of a medullary centre. The optical reflex is slower; in man, the nervous pathway includes the visual cortex (the outer substance of the brain; the visual centre is located in the occipital—rear—lobe). The reflex is absent in children of less than nine months.

Normal rhythm. In the waking hours the eyes blink fairly regularly at intervals of two to 10 seconds, the actual rate being a characteristic of the individual. The function of this is to spread the lacrimal secretions over the cornea. It might be thought that each blink would be reflexly determined by a corneal stimulus—drying and irritation—but extensive studies indicate that this view is wrong; the normal blinking rate is apparently determined by the activity of a "blinking centre" in the globus pallidus of the caudate nucleus, a mass of nerve cells between the base and the outer substance of the brain. This is not to deny that the blink rate is modified by external stimuli.

There is a strong association between blinking and the action of the extraocular muscles. Eye movement is generally accompanied by a blink, and it is thought that this aids the eyes in changing their fixation point.

Secretion of tears. The exposed surface of the globe (eyeball) is kept moist by the tears secreted by the lacrimal apparatus, together with the mucous and oily secretions of the other secretory organs and cells of the lids and conjunctiva; these have been described earlier. The secretion produces what has been called the precorneal film, which consists of an inner layer of mucus, a middle layer of lacrimal secretion, and an outer oily film that reduces the rate of evaporation of the underlying watery layer. The normal daily (24-hour) rate of secretion has been estimated at about 0.75 to 1.1 grams (0.03–0.04 ounce avoirdupois); secretion tends to decrease with age. Chemical analysis of the tears reveals a typical body fluid with a salt concentration similar to that of blood plasma. An interesting component is lysozyme, an enzyme that has bactericidal action by virtue of its power of dissolving away the outer coats of many bacteria.

Tear reflexes

Tears are secreted reflexly in response to a variety of stimuli—*e.g.*, irritative stimuli to the cornea, conjunctiva, nasal mucosa; hot or peppery stimuli applied to the mouth and tongue; or bright lights. In addition, tear flow occurs in association with vomiting, coughing, and yawning. The secretion associated with emotional upset is called psychical weeping. Severing of the sensory root of the trigeminal

(fifth cranial) nerve prevents all reflex weeping, leaving psychical weeping unaffected; similarly, the application of cocaine to the surface of the eye, which paralyzes the sensory nerve endings, inhibits reflex weeping, even when the eye is exposed to potent tear gases. The afferent (sensory) pathway in the reflex is thus by way of the fifth cranial, or the trigeminal nerve. The motor innervation is by way of the autonomic (involuntary) division; the parasympathetic supply derived from the facial nerve (the seventh cranial nerve) seems to have the dominant motor influence. Thus, drugs that mimic the parasympathetic, such as acetylcholine, provoke secretion, and secretion may be blocked by such typical anticholinergic drugs as atropine. Innervation of the lacrimal gland is not always complete at birth, so that the newborn infant is generally said to cry without weeping. Because absence of reflex tearing fails to produce any serious drying of the cornea, and surgical destruction of the main lacrimal gland is often without serious consequences, it seems likely that the subsidiary secretion from the accessory lacrimal glands is adequate to keep the cornea moist. The reflex secretion that produces abundant tears may be regarded as an emergency response.

A drainage mechanism for tears is necessary only during copious secretion. The mechanism, described as the lacrimal pump, consists of alternately negative and positive pressure in the lacrimal sac caused by the contraction of the orbicularis muscle during blinking.

Movements of the eyes. Because only a small portion of the retina, the fovea, is actually employed for distinct vision, it is vitally important that the motor apparatus governing the direction of gaze be extremely precise in its operation, and rapid. Thus, the gaze must shift swiftly and accurately during the process of reading. Again, if the gaze must remain fixed on a single small object—*e.g.*, a golf ball—the eyes must keep adjusting their gaze to compensate for the continuous small movements of the head and to maintain the image exactly on the fovea. The extraocular muscles that carry out these movements are under voluntary control; thus, the direction of regard can be changed deliberately. Most of the actual movements of the eyes are carried out without awareness, however, in response to movements of the objects in the environment, or in response to movements of the head or the rest of the body, and so on. In examining the mechanisms of the eye movements, then, one must resolve them into a number of reflex responses to changes in the environment or the individual, remembering, of course, that there is an overriding voluntary control.

Centring images on the fovea

The axes of the eye. It is worthwhile at this point to define certain axes of the eyes employed during different types of study. The optic axis of the eye is a line drawn through the centre of the cornea and the nodal (central) point of the eye; it actually does not intersect with the retina at the centre of the fovea as might be expected, but toward the nose from this, so that there is an angle of about five degrees between (1) the visual axis—the line joining the point fixated (the point toward which the gaze is directed) and the nodal point—and (2) the optic axis.

Actions of muscles. The general modes of action of the six extraocular muscles have been described in connection with their anatomy: rotation of the eye toward the nose is carried out by the medial rectus; outward movement is by the lateral rectus. Upward movements are carried out by the combined actions of the superior rectus and the inferior oblique muscles, and downward movements by the inferior rectus and the superior oblique. Intermediate directions of gaze are achieved by combined actions of several muscles. When the two eyes act together, as they normally do, and change their direction of gaze to the left, for example, the left eye rotates away from the nose by means of its lateral rectus, while the right eye turns toward the nose by means of its medial rectus. These muscles may be considered as a linked pair; that is, when they are activated by the central nervous system this occurs conjointly and virtually automatically. This linking of the muscles of the two eyes is an important physiological feature and has still more important pathological interest in the analysis of squint, when the two eyes fail to be directed at the same point.

Binocular movements. The binocular movements (the movements of the two eyes) fall into two classes, the conjugate movements, when both eyes move in the same direction, as in a change in the direction of gaze, and disjunctive movements, when the eyes move in opposite directions. Thus, during convergence onto a near object both eyes move toward the nose; the movement is horizontal, but disjunctive, by contrast with the conjugate movement when both eyes move, say, to the right. The disjunctive movement of convergence can be carried out voluntarily, but the act is usually brought about reflexly in response to the changed optical situation—*i.e.*, the nearness of the object of gaze. A seesaw movement of the eyes, whereby one eye looks upward and the other downward, is possible, but not voluntarily; to achieve this a prism is placed in front on one eye so that the object seen through it appears displaced upward or downward; the other eye sees the object where it is. The result of such an arrangement is that, unless the eye with the prism in front makes an upward or downward movement, independent of the other, the images will not fall on corresponding parts of the retinas in the two eyes. Such a noncorrespondence of the retinal images causes double vision; to avoid this, there is an adjustment in the alignment of the eyes so that a seesaw movement is actually executed. In a similar way, the eyes may be made to undergo torsion, or rolling. A conjugate torsion, in which both eyes rotate about their anteroposterior (fore-and-aft) axes in the same sense, occurs naturally; for example, when the head tips toward one shoulder the eyes tend to roll in the opposite direction, with the result that the image of the visual field on the retina tends to remain vertical in spite of the rotation of the head.

Conjugate and disjunctive movements

Nervous control. The nerves controlling the actions of the muscles are the third, fourth, and sixth cranial nerves, with their bodies (nuclei) in the brainstem; the third, or oculomotor nerve, controls the superior and inferior recti, the medial rectus, and inferior oblique; the fourth cranial nerve, the trochlear nerve, controls the superior oblique; and the sixth, the abducens nerve, controls the lateral rectus. The nuclei of these nerves are closely associated; especially, there are connections between the nuclei of the sixth cranial nerve, controlling the lateral rectus, and the nucleus of the third, controlling the medial rectus; it is through this close relationship that the linking of the lateral rectus of one eye and the medial rectus of the other, indicated above, is achieved. Another type of linking is concerned with reciprocal inhibition; that is, when there are two antagonistic muscles, such as the medial and the lateral rectus, contraction of one is accompanied by a simultaneous inhibition of the other. Muscles show a continuous slight activity even when at rest; this keeps them taut; this action, called tonic activity, is brought about by discharges in the motor nerve to the muscle. Hence, when the agonist muscle contracts its antagonist must be inhibited.

Reflex pathways. In examining any reflex movement one must look for the sensory input—*i.e.*, the way in which messages in sensory nerves bring about discharges in the motor nerves to the muscles; this study involves the connections of the motor nerves or nuclei with other centres of the brain.

When a subject is looking straight ahead and a bright light appears in the periphery of his field of vision, his eyes automatically turn to fix on the light; this is called the fixation reflex. The sensory pathway in the reflex arc leads as far as the cerebral cortex because removal of the occipital cortex (the outer brain substance at the back of the head) abolishes reflex eye movements in response to light stimuli. If the occipital cortex is stimulated electrically, movements of the eyes may be induced, and in fact one may draw a pattern of the visual field on the occipital cortex corresponding with the directions in which the gaze is turned when given points on the cortex are stimulated. This pattern corresponds with the pattern obtained by recording the visual responses to light stimuli from different parts of the visual field.

Fixation reflex

The remainder of the pathway—*i.e.*, from the occipital cortex to the motor neurons in the brainstem—has long

been considered to involve the superior colliculi as relay stations, and they certainly have such a role in lower animals; but in human beings a pathway from the cortex to the eye-muscle nuclei independent of the superior colliculi of the midbrain is now generally assumed.

Continual movements of the eyes occur even when an effort is made to maintain steady fixation of an object. Some of these movements may be regarded as manifestations of the fixation reflex; thus, the eyes tend to drift off their target, and, because of this, the fixation reflex comes into play, bringing the eyes back on target.

Nystagmus

Experimentally, the fixation reflex can be studied by observation of the regular to-and-fro movements of the eyes as they follow a rotating drum striped in black and white. (Such movements of the eyes directed at a moving object are called optokinetic nystagmus; nystagmus itself is the involuntary movement of the eye back and forth, up and down, or in a rotatory or a mixed fashion.) While the eyes watch the moving drum, they involuntarily make a slow movement as a result of fixing their gaze on a particular stripe. At a certain point, fixation is broken off, and the eyes spring back to fix on a new stripe. Thus, the nystagmus consists of a slow movement with angular velocity equal to that of the rotation of the drum, then a fast saccade, or jump from one point of fixation to another, in the opposite direction; the process is repeated indefinitely.

Another type of nystagmus reveals the play of another set of reflexes. These are mediated by the semicircular canals—*i.e.*, the organs of balance or the vestibular apparatus. Such a reflex may be evoked by rotating the subject in a chair at a steady speed; the eyes move slowly in the opposite direction to that of rotation and, at the end of their excursion, jump back with a fast saccade in the direction of rotation. If rotation suddenly ceases, the eyes go into a nystagmus in the opposite direction, the postrotatory nystagmus.

During rotation, certain semicircular canals are being stimulated, and the important point is that any acceleration of the head that stimulates these canals will cause reflex movements of the eyes; thus, acceleration of the head to the right causes a movement of the eyes to the left, the function of the reflex being to enable the eyes to maintain steady fixation of an object despite movements of the head. The reflex occurs even when the eyes are shut, and, when the eyes are open, it obviously cooperates with the fixation reflex in maintaining steady fixation. In many lower animals this connection between organs of balance and eyes is very rigid; thus, one may move the tail of a fish, and its eyes will move reflexly. In man, not only do the semicircular canals function in close relation to the eye muscles but so also do the gravity organ—the utricle—and the stretch receptors in the muscles of the neck. Thus, when the head is turned upward, there is a reflex tendency for the eyes to move downward, even if the eyes are shut. The actual movement is probably initiated by the reflex from the semicircular canals, which respond to acceleration, but the maintenance of the position is brought about by a reflex through the stretch of the neck muscles and also through the pull of gravity on the utricle, or otolith organ, in the inner ear.

Voluntary centre. The eyes are under voluntary control, and it is thought that the cortical area subserving voluntary eye movements is in the frontal cortex. Stimulation of this in primates causes movements of the eyes that are well coordinated, and a movement induced by this region prevails over one induced by stimulation of the occipital cortex. The existence of a separate centre in man is revealed by certain neurological disorders in which the subject is unable to fixate voluntarily but can do so reflexly; *i.e.*, he can follow a moving light.

The nature of eye movements. So far, the relation of the movements of the eyes to the requirements of the visual apparatus and their control have been touched upon. To examine the character of the movements in some detail requires rapid, accurate measurement of the movements that the eyes undergo. Modern studies of this subject employ a contact lens fitting on to the globe; on the lens is a small plane mirror, and a parallel bundle of rays is reflected off this mirror onto a moving film.

By the use of refined methods of measuring the position of the eyes at any moment, it becomes immediately evident that the eyes are never stationary for more than a fraction of a second; the movements are of three types: (1) irregular movements of high frequency (30–70 per second) and small excursions of about 20 seconds of arc; (2) flicks, or saccades, of several minutes of arc occurring at regular intervals of about one second; and between these saccades there occur (3) slow irregular drifts extending up to six minutes of arc. The saccades are corrective, serving to bring the fixation axis on the point of regard after this has drifted away from it too far, and thus are a manifestation of the fixation reflex.

The significance of these small movements during fixation was revealed by studies on the stabilized retinal image: by a suitable optical device the image of an object could be held stationary on the retina in spite of the movements of the eye. It was found that under these conditions the image would disappear within a few seconds. Thus, the movements of the eye are apparently necessary to allow the contours of the image to fall on a new set of rods and cones at repeated intervals; if this does not occur, the retina adapts to their stimulus and ceases to send messages to the central nervous system. The small flicks mentioned above are essentially the same as the larger movement made when the two eyes fixate (fix on) a light when it suddenly appears in the peripheral field; this is given the general name of the saccade, to distinguish it from the slower movements occurring during convergence and smooth following. The dynamics of the saccade have been studied in some detail by several workers. There is a reaction time of about 120 to 180 milliseconds, after which both eyes move simultaneously; there is a definite overshoot and, with an excursion of 20°, the operation is completed in about 90 milliseconds. The maximum velocity increases with the extent of the movement, being 300° per second for 10° and 500° per second for 30°. A remarkable feature is the apparent absence of significant inertia in the eyeball, so that movement is halted, not by any checking action of antagonistic muscles but simply by cessation of contraction of the agonists; thus, the movement is not ballistic. Once under way, the saccade is determined in amount, so that the subject cannot voluntarily alter its direction and extent. The control mechanism for the saccadic type of movement can be described as a sampled data system, *i.e.*, the brain makes discontinuous samples of the position of the eyes in relation to the target and corrects the error, in contrast to a continuous feedback system that takes account of the error all the time.

The saccade, disjunctive movements, and tracking

The movements of the eyes when they converge onto a near object are in remarkable contrast to the saccade; the angular velocity is only about 25° per second, compared with values as high as 500° per second in the saccade. The great difference in speed suggested to two investigators that the two movements are executed by different muscle fibres. In fact, the extraocular muscles do contain two types of muscle fibre with characteristically different nerve supplies, and some recent studies tend to support this view of a dual mechanism.

If a moving light suddenly appears in the field of view, and if its rate of movement is less than about 30° per second, the response of the eyes is remarkably efficient; a saccade brings the eyes on target, and they follow the motion at almost exactly the same angular velocity as that of the target; inaccuracies in following lead to corrective saccades. When the rate of movement of the target is greater than about 30° per second, these corrective saccades become more obvious because now smooth following is not possible; the eyes make constant-velocity movements, but the velocity rarely matches that of the moving target, so that there must be frequent corrective saccades. Studies have shown that the following movements are highly integrated and must involve a continuous feedback system whereby errors are used to modify the performance. Thus, the systems for control of saccades and tracking movements are fundamentally different.

Vision suppression during a saccade. If one looks into a mirror and fixates one of one's eyes and then fixates the other, one does not see the eyes moving; and it has

been argued that, during an eye movement, vision is suppressed; if vision were not suppressed, moreover, it seems likely that the images of the external world would appear smeared during a movement. Experimental studies have shown that there is, indeed, a suppression of vision during a saccade.

THE WORK OF THE OPTICAL LENS SYSTEM

Refraction by cornea and lens. The optical system of the eye is such as to produce a reduced inverted image of the visual field on the retina; the system behaves as a convex lens but is, in fact, much more complex, refraction taking place not at two surfaces, as in a lens, but at four separate surfaces—at the anterior and the posterior surfaces of the cornea and of the crystalline lens. Each of these surfaces is approximately spherical, and at each optical interface—*e.g.*, between air and the anterior surface of the cornea—the bending of a ray of light is toward the axis, so that, in effect, there are four surfaces tending to make rays of light converge on each other. If the rays of light falling on the cornea are parallel—*i.e.*, if they come from a distant point—the net effect of this series of refractions at the four surfaces is to bring these rays to a point focus of the optical system, which in the normal, or emmetropic, eye corresponds with the retina. The greatest change of direction, or bending of the rays, occurs where the difference of refractive index is greatest, and this is when light passes from air into the cornea, the refractive index of the corneal substance being 1.3376; the refractive indices of the cornea and aqueous humour are not greatly different, that of the aqueous humour being 1.336 (as is that of the vitreous); thus, the bending, as the rays meet the concave posterior surface of the cornea and emerge into a medium of slightly less refractive index, is small. The lens has a greater refractive index than that of its surrounding aqueous humour and vitreous body, 1.386 to 1.406, so that its two surfaces contribute to convergence, the posterior surface normally more than the anterior surface because of its greater curvature (smaller radius).

Four refractive surfaces

Normal sightedness and near- and farsightedness. In contrast to the focussing of the normal (emmetropic) eye, in which the image of the visual field is focussed on the retina, the image may be focussed in front of the retina (nearsightedness, or myopia), or behind the retina (farsightedness or hyperopia). In myopia the vision of distant objects is not distinct because the image of a distant point falls within the vitreous and the rays spread out to form a blur circle on the retina instead of a point. In this condition the eye is said to have too great dioptric (refractive) power for its length. When the focus falls behind the retina, the image of the distant point is again a circle on the retina; and the farsighted eye is said to have too little dioptric power. The important point to appreciate is that emmetropia, or normal sight, requires that the focal power of the dioptric system be matched to the axial length of the eye; it certainly is remarkable that emmetropia is indeed the most common condition when it is appreciated that just one millimetre of error in the matching of axial length with focal length would cause a person to require a spectacle correction. In general, however, the effects of variations in dimensions tend to compensate each other. Thus, for example, an unusually large eye might, at first thought, be expected to be myopic, but a large eye tends to be associated with a large radius of curvature of the cornea, and this would reduce the power—*i.e.*, increase the focal length—and so an unusually large eye is not necessarily a myopic one.

Accommodation. *Effects of accommodation.* The image of an object brought close to the eye would be formed behind the retina if there were no change in the focal length of the eye. This change to bring the image of an object upon the retina is called accommodation. The point nearer than which accommodation is no longer effective is called the near point of accommodation. In very young people, the near point of accommodation is quite close to the eye, namely about seven centimetres (about three inches) in front at 10 years old; at 40 years the distance has increased to about 16 centimetres (about 6 inches), and at 60 years it is 100 centimetres or one metre (39 inches). Thus, a 60-year-old would not be able to read a book held at the convenient distance of about 40 centimetres (16 inches), and the extra power required would have to be provided by convex lenses in front of the eye, an arrangement called the presbyopic correction.

The near point of accommodation

Mechanism of accommodation. It is essentially an increase in curvature of the anterior surface of the lens that is responsible for the increase in power involved in the process of accommodation. A clue to the way in which this change in shape takes place is given by the observation that a lens that has been taken out of the eye is much rounder and fatter than one within the eye; thus, its attachments by the zonular fibres to the ciliary muscle within the eye preserve the unaccommodated or flattened state of the lens; and modern investigations leave little doubt that it is the pull of the zonular fibres on the elastic capsule of the lens that holds the anterior surface relatively flat. When these zonular fibres are loosened, the elastic tension in the capsule comes into play and remolds the lens, making it smaller and thicker. Thus, the physiological problem is to find what loosens the zonular fibres during accommodation. The ciliary muscle has been described earlier, and it has been shown that the effect of contracting its fibres is, in general, to pull the whole ciliary body forward and to move the anterior region toward the axis of the eye by virtue of the sphincter action of the circular fibres. Both of these actions will slacken the zonular fibres and therefore allow the change in shape. As to why it is the anterior surface that changes most is not absolutely clear, but it is probably a characteristic of the capsule rather than of the underlying lens tissue. Defective accommodation in presbyopia is not due to a failure of the ciliary muscle but rather to a hardening of the substance of the lens with age to the point that readjustments of its shape become ever more difficult.

Nerve action. Accommodation is an involuntary reflex act, and the ciliary muscle belongs to the smooth involuntary class. Appropriate to this, the innervation is through the autonomic system, the parasympathetic nerve cells belonging to the oculomotor nerve (the third cranial nerve) occupying a special region of the nucleus in the midbrain called the Edinger-Westphal nucleus; the fibres have a relay point in the ciliary ganglion in the eye socket, and the postganglionic fibres enter the eye as the short ciliary nerves. The stimulus for accommodation is the nearness of the object, but the manner in which this nearness is translated into a stimulus is not clear. Thus, the fact that the image is blurred is not sufficient to induce accommodation; the eye has some power of discriminating whether the blurredness is due to an object being too far away or too close, so that something more than mere blurredness is required.

The pupil. The amount of light entering the eye is restricted by the aperture in the iris, the pupil.

The light and near reflexes

When a person is in a dark room his pupil is large, perhaps eight millimetres (0.3 inch) in diameter, or more. When the room is lighted there is an immediate constriction of the pupil, the light reflex; this is bilateral, so that even if only one eye is exposed to the light both pupils contract to nearly the same extent. After a time the pupils expand even though the bright light is maintained, but the expansion is not large. The final state is determined by the actual degree of illumination; if this is high, then the final state may be a diameter of only about three to four millimetres (about 0.15 inch); if it is not so high, then the initial constriction may be nearly the same, but the final state may be with a pupil of four to five millimetres (about 0.18 inch). During this steady condition, the pupils do not remain at exactly constant size; there is a characteristic oscillation in size that, if exaggerated, is called hippus.

A pupillary constriction will also occur when a person looks at a near object—the near reflex. Thus, accommodation and pupillary constriction occur together reflexly and are excited by the same stimulus. The function of the pupil is clearly that of controlling the amount of light entering the eye, and hence the light reflex. The constriction occurring during near vision suggests other functions, too; thus, the aberrations of the eye (failure of some refracted rays to focus on the retina) are decreased by reducing the

aperture of its optical system. In the dark, aberrations are of negligible significance, so that a person is concerned only with allowing as much light into the eye as possible; in bright light high visual acuity is usually required, and this means reducing the aberrations. The depth of focus of the optical system is increased when the aperture is reduced, and the near reflex is probably concerned with increasing depth of focus under these conditions.

Dilation of the pupil occurs as a result of strong psychical stimuli and also when any sensory nerve is stimulated; dilation thus occurs in extreme fear and in pain.

Neuromuscular mechanisms. The muscles of the iris have been described earlier. It is clear from their general features that constriction of the pupil is brought about by shortening of the circular ring of fibres—the sphincter; dilation is brought about by shortening of the radially oriented fibres. The sphincter is innervated by parasympathetic fibres of the oculomotor nerve, with their cell bodies in the Edinger-Westphal nucleus, as are the nerve cells controlling accommodation; thus, the close association between the accommodation and pupillary reflexes is reflected in a close anatomical contiguity of their motor nerve cells.

The sensory pathway in the light reflex involves the rods and cones, bipolar cells, and ganglion cells. As indicated earlier, a relay centre for pupillary responses to light is the pretectal nucleus in the midbrain. There is a partial crossing-over of the fibres of the pretectal nerve cells so that some may run to the motor nerve cells in the Edinger-Westphal nucleus of both sides of the brain, and it is by this means that illumination of one eye affects the other. The Edinger-Westphal motor neurons have a relay point in the ciliary ganglion, a group of nerve cells in the eye socket, so that its electrical stimulation causes both accommodation and pupillary constriction; similarly, application of a drug, such as pilocarpine, to the cornea will cause a constriction of the pupil and also a spasm of accommodation; atropine, by paralyzing the nerve supply, causes dilation of the pupil and paralysis of accommodation (cycloplegia).

Dilator response

The dilator muscle of the iris is activated by sympathetic nerve fibres. Stimulation of the sympathetic nerve in the neck causes a powerful dilation of the iris; again, the influx of adrenalin into the blood from the adrenal glands during extreme excitement results in pupillary dilation.

Many involuntary muscles receive a double innervation, being activated by one type of nerve supply and inhibited by the other; modern experimentation indicates that the iris muscles are no exception, so that the sphincter has an inhibitory sympathetic nerve supply, while the dilator has a parasympathetic (cholinergic) inhibitor. Thus, a drug like pilocarpine not only activates the constrictor muscle but actively inhibits the dilator. A similar double innervation has been described for the ciliary muscle. In general, any change in pupillary size results from a reciprocal innervation of dilator and constrictor; thus, activation of the constrictor is associated with inhibition of the dilator and vice versa.

The near response. In general, as has been indicated, pupillary constriction and accommodation occur together, in response to the same stimulus; a third element in this near response is, of course, the convergence (turning in) of the eyes, mediated by voluntary muscles, the medial recti. Experimentally, it is often possible to separate these activities, in the sense that one may cause convergence without accommodation by placing appropriate prisms in front of the eyes; or one may cause accommodation without convergence by placing diverging lenses in front of the eyes. There are many experiments that show that accommodation and convergence are neurologically linked to some extent, however.

THE WORK OF THE RETINA

Some basic facts of vision. So far, attention has been directed to what are essentially the preliminaries to vision; it is now time to examine some of the elementary facts of vision and to relate them to the structure of the retina and, later, to chemically identifiable events.

Measurement of the threshold. An important means of measuring a sensation is to determine the threshold stimulus—*i.e.*, the minimum energy required to evoke the sensation. In the case of vision, this would be the minimum number of quanta of light entering the eye in unit time. If it is found that the threshold has altered because of a change of some sort, then this change can be said to have altered the subject's sensitivity to light, and a numerical value can be assigned to the sensitivity by use of the reciprocal of the threshold energy. Practically, a subject may be placed in the dark in front of a white screen, and the screen may be illuminated by flashes of light; for any given intensity of illumination of the screen, it is not difficult to calculate the flow of light energy entering the eye. One may begin with a low intensity of flash and increase this successively until the subject reports that he can see the flash. In fact, at this threshold level, he will not see every flash presented, even though the intensity of the light is kept constant; for this reason, a certain frequency of seeing—*e.g.*, four times out of six—must be selected as the arbitrary point at which to fix the threshold.

Decline in visual threshold

When measurements of this sort are carried out, it is found that the threshold falls progressively as the subject is maintained in the dark room. This is not due to dilation of the pupil because the same phenomenon occurs if the subject is made to look through an artificial pupil of fixed diameter. The eye, after about 30 minutes in the dark, may become about 10,000 times more sensitive to light. Vision under these conditions is, moreover, characteristically different from what it is under ordinary daylight conditions. Thus, in order to obtain best vision, the eye must look away from the screen so that the image of the screen does not fall on the fovea; if the screen is continuously illuminated at around this threshold level it will be found to disappear if its image is brought onto the fovea, and it will become immediately visible on looking away. The same phenomenon may be demonstrated on a moonless night if the gaze is fixed on a dim star; it disappears on fixation and reappears on looking away. This feature of vision under these near-threshold or scotopic conditions suggests that the cones are effectively blind to weak light stimuli, since they are the only receptors in the fovea. This is the basis of the duplicity theory of vision, which postulates that when the light stimulus is weak and the eye has been dark-adapted, it is the rods that are utilized because, under these conditions, their threshold is much lower than that of the cones. When the subject first enters the dark, the rods are the less sensitive type of receptor, and the threshold stimulus is the light energy required to stimulate the cones; during the first five or more minutes the threshold of the cones decreases; *i.e.*, they become more sensitive. The rods then increase their sensitivity to the point that they are the more sensitive, and it is they that now determine the sensitivity of the whole eye, the threshold stimuli obtained after 10 minutes in the dark, for example, being too weak to activate the cones.

Duplicity theory of vision

Scotopic sensitivity curve. When different wavelengths of light are employed for measuring the threshold, it is found, for example, that the eye is much more sensitive to blue-green light than to orange. The interesting feature of this kind of study is that the subject reports only that the light is light; he distinguishes no colour. If the intensity of a given wavelength of light is increased step by step above the threshold, a point comes when the subject states that it is coloured, and the difference between the threshold for light appreciation and this, the chromatic threshold, is called the photochromatic interval. This suggests that the rods give only achromatic, or colourless, vision, and that it is the cones that permit wavelength discrimination. The photochromatic interval for long wavelengths (red light) is about zero, which means that the intensity required to reach the sensation of light is the same as that to reach the sensation of colour. This is because the rods are so insensitive to red light; if the dark-adaptation curve is plotted for a red stimulus it is found that it follows the cone path, like that for foveal vision at all wavelengths.

Loss of dark adaptation. If, when the subject has become completely dark-adapted, one eye is held shut and the other exposed to a bright light for a little while, it is found that, whereas the dark-adapted eye retains its high

sensitivity, that of the light-exposed eye has decreased greatly; it requires another period of dark adaptation for the two eyes to become equally sensitive.

These simple experiments pose several problems, the answers to which throw a great deal of light on the whole mechanism of vision. Why, for example, does it require time for both rods and cones to reach their maximum sensitivity in the dark? Again, why is visual acuity so low under scotopic conditions compared with that in daylight, although sensitivity to light is so high? Finally, why do the rods not serve to discriminate different wavelengths?

Bleaching of rhodopsin. It may be assumed that a receptor is sensitive to light because it contains a substance that absorbs light and converts this vibrational type of energy into some other form that is eventually transmuted into electrical changes, and that these may be transmitted from the receptor to the bipolar cell with which it is immediately connected. When the retina of a dark-adapted animal is removed and submitted to extraction procedures, a pigment, originally called visual purple but now called rhodopsin, may be obtained. If the eye is exposed to a bright light for some time before extraction, little or no rhodopsin is obtained. When retinas from animals that had been progressively dark-adapted were studied, a gradual increase in the amount of rhodopsin that could be extracted was observed. Thus, rhodopsin, on absorption of light energy, is changed to some other compound, but new rhodopsin is formed, or rhodopsin is regenerated, during dark adaptation. The obvious inference is that rhodopsin is the visual pigment of the rods, and that when it is exposed to relatively intense lights it becomes useless for vision. When the eye is allowed to remain in the dark the rhodopsin regenerates and thus becomes available for vision. There is now conclusive proof that rhodopsin is, indeed, the visual pigment for the rods; it is obtained from retinas that have only rods and no cones—*e.g.*, the retinas of the rat or guinea pig, and it is not obtained from the pure cone retina of the chicken.

Rhodopsin as the photopigment

When the absorption spectrum is measured, it is found that its maximum absorption occurs at the point of maximum sensitivity of the dark-adapted eye. Similar measurements may be carried out on animals, but the threshold sensitivity must be determined by some objective means—*e.g.*, the response of the pupil, or, better still, the electrical changes occurring in the retina in response to light stimuli. Thus, the electroretinogram (ERG) is the record of changes in potential between an electrode placed on the surface of the cornea and an electrode placed on another part of the body, caused by illumination of the eye.

The high sensitivity of the rods by comparison with the cones may be a reflection of the greater concentration in them of pigment that would permit them to catch light more efficiently, or it may depend on other factors—*e.g.*, the efficiency of transformation of the light energy into electrical energy. The pigments responsible for cone vision are not easily extracted or identified, and the problem will be considered in the material on colour vision. An important factor, so far as sensitivity is concerned, is the actual organization of the receptors and neurons in the retina.

Synaptic organization of the retina. The basic structure of the retina has been indicated earlier. As in other parts of the nervous system, the messages initiated in one element are transmitted, or relayed, to others. The regions of transmission from one cell to another are areas of intimate contact known as synapses. An impulse conveyed from one cell to another travels from the first cell body along a projection called an axon, to a synapse, where the impulse is received by a projection, called a dendrite, of the second cell. The impulse is then conveyed to the second cell body, to be transmitted further, along the second cell's axon.

It will be recalled that the functioning cells of the retina are the receptor cells—the rods and cones; the ganglion cells, the axons of which form the optic nerve; and cells that act in a variety of ways as intermediaries between the receptors and the ganglion cells. These intermediaries are named bipolar cells, horizontal cells, and amacrine cells.

Plexiform layers. As was indicated earlier, the synapses occur in definite layers, the outer and inner plexiform layers. In the outer plexiform layer the bipolar cells make their contacts, by way of their dendrites, with the rods and cones, specifically the spherules of the rods and the pedicles of the cones. In this layer, too, the projections from horizontal cells make contacts with rods, cones, and bipolar cells, giving rise to a horizontal transmission and thereby allowing activity in one part of the retina to influence the behaviour of a neighbouring part. In the inner plexiform layer, the axons of the bipolar cells make connection with the dendrites of ganglion cells, once again at special synaptic regions. (The dendrites of a nerve cell carry impulses to the nerve cell; its axon, away from the cell.) Here, too, a horizontal interconnection between bipolar cells is brought about, in this case by way of the axons and dendrites of amacrine cells.

Inner plexiform layer

The bipolar cells are of two main types: namely, those that apparently make connection with only one receptor—a cone—and those that connect to several receptors. The type of bipolar cell that connects to a single cone is called the midget bipolar. The other type of bipolar cell is called diffuse; varieties of these include the rod bipolar, the dendritic projections of which spread over an area wide enough to allow contacts with as many as 50 rods; and the flat cone bipolar, which collects messages from up to seven cones.

Ganglion cells are of two main types: namely, the midget ganglion cell, which apparently makes a unique connection with a midget bipolar cell, which in turn is directly connected to a single cone; and a diffuse type, which collects messages from groups of bipolar cells.

Convergence of the messages. The presence of diffuse bipolar and ganglion cells collecting messages from groups of receptors and bipolar cells, and, what may be even more important, the presence of lateral connections of groups of receptors and bipolar cells through the horizontal and amacrine cells, means that messages from receptors over a rather large area of the retina may converge on a single ganglion cell. This convergence means that the effects of light falling on the receptive field may be cumulative, so that a weak light stimulus spread over about 1,000 rods is just as effective as a stronger stimulus spread over 100 or less; in other words, a large receptive field will have a lower threshold than a small one; and this is, in fact, the basis for the high sensitivity of the area immediately outside the fovea, where there is a high density of rods that converge on single bipolar cells. Thus, if it is postulated that the cones do not converge to anything like the same extent as the rods, the greater sensitivity of the latter may be explained; and the anatomical evidence favours this postulate.

It has been indicated above that the regeneration of visual pigment is a cause of the increased sensitivity of the rods that occurs during dark adaptation. This, apparently, is only part of the story. An important additional factor is the change in functional organization of the retina during adaptation. When the eye is light-adapted, functional convergence is small, and sensitivity of rods and cones is low; as dark adaptation proceeds, convergence of rods increases. The anatomical connections do not change, but the power of the bipolar cells and ganglion cells to collect impulses is increased, perhaps by the removal of an inhibition that prevents this during high illumination of the retina.

Absolute threshold and minimum stimulus for vision. As was indicated earlier, the threshold is best indicated in terms of frequency of seeing since, because of fluctuations in the threshold, there is no definite luminance of a test screen at which it is always seen by the observer, and there is no luminance just below this at which it is never seen. Experiments, in which 60 percent was arbitrarily taken as the frequency of seeing and in which the image of a patch of light covered an area of retina containing about 20,000,000 rods, led to the calculation that the mean threshold stimulus represents 2,500 quanta of light that is actually absorbed per square centimetre of retina. This calculation leads to two important conclusions: namely, that at the threshold only one rod out of thousands comes into operation, and that during the application of a short stimulus the chances are that no rod receives more than a single quantum.

A quantum, defined as the product of Planck's constant (6.63×10^{-27} erg-second) times the frequency of light, is the minimum amount of light energy that can be employed. A rod excited by a single quantum cannot excite a bipolar cell without the simultaneous assistance of one or more other rods. Experiments carried out in the 1940s indicated that a stimulus of about 11 quanta is required; thus it may require 11 excited rods, each receiving one quantum of light, to produce the sensation of light.

Uncertainty principle

Quantum fluctuations. With such small amounts of energy as those involved in the threshold stimulus, the uncertainty principle becomes important; according to this, there is no certainty that a given flash will have the expected number of quanta in it, but only a probability. Thus, one may speak of a certain average number of quanta and the actual number in any given flash, and one may compute on statistical grounds the shape of curve that is obtained by plotting frequency with which a flash contains, say, four quanta or more against the average number in the flash. One may also plot the frequency with which a flash is seen against the average number of quanta in the flash, and this frequency-of-seeing curve turns out to be similar to the frequency-of-containing-quanta curve when the number of quanta chosen is five to seven, depending on the observer. This congruence strongly suggests that the fluctuations in response to a flash of the same average intensity are caused by fluctuations in the energy content of the stimulus, and not by fluctuations in the sensitivity of the retina.

Spatial summation. In spatial summation two stimuli falling on nearby areas of the retina add their effects so that either alone may be inadequate to evoke the sensation of light, but, when presented simultaneously, they may do so. Thus, the threshold luminance of a test patch required to be just visible depends, within limits, on its size, a larger patch requiring a lower luminance, and vice versa. Within a small range of limiting area, namely that subtending about 10 to 15 minutes of arc, the relationship called Ricco's law holds; *i.e.,* threshold intensity multiplied by the area equals a constant. This means that over this area, which embraces several hundreds of rods, light falling on the individual rods summates, or accumulates, its effects completely so that 100 quanta falling on a single rod are as effective as one quantum falling simultaneously on 100 rods. The basis for this summation is clearly the convergence of receptors on ganglion cells, the chemical effects of the quanta of light falling on individual rods being converted into electrical changes that converge on a single bipolar cell through its branching dendritic processes. Again, the electrical effects induced in the bipolar cells may summate at the dendritic processes of a ganglion cell so that the receptive field of a ganglion cell may embrace many thousands of rods.

Temporal summation. In temporal summation, two stimuli, each being too weak to excite, cause a sensation of light if presented in rapid succession on the same spot of the retina; thus, over a certain range of times, up to 0.1 second, the Bunsen-Roscoe law holds: namely, that the intensity of light multiplied by the time of exposure equals a constant. Thus it was found that within this time interval (up to 0.1 second), the total number of quanta required to excite vision was 130, irrespective of the manner in which these were supplied. Beyond this time, summation was still evident, but it was not perfect, so that if the duration was increased to one second the total number of quanta required was 220. Temporal summation is consistent with quantum theory; it has been shown that fluctuations in the number of quanta actually in a light flash are responsible for the variable responsiveness of the eye; increasing the duration of a light stimulus increases the probability that it will contain a given number of quanta, and that it will excite.

Inhibition. In the central nervous system generally, the relay of impulses from one nerve cell or neuron to excite another is only one aspect of neuronal interaction. Just as important, if not more so, is the inhibition of one neuron by the discharge in another. So it is in the retina. Subjectively, the inhibitory activity is reflected in many of the phenomena associated with adaptation to light or its reverse. Thus, the decrease in sensitivity of the retina to light during exposure to light is only partially accounted for by bleaching of visual pigment, be it the pigment in rod or cone; an important factor is the onset of inhibitory processes that reduce the convergence of receptors on ganglion cells. Some of the rapidly occurring changes in sensitivity described as alpha adaptation are doubtless purely neural in origin.

Many so-called inductive phenomena indicate inhibitory processes; thus, the phenomenon of simultaneous contrast, whereby a patch of light appears much darker if surrounded by a bright background than by a black, is due to the inhibitory effect of the surrounding retina on the central region, induced by the bright surrounding. Many colour-contrast phenomena are similarly caused; thus, if a blue light is projected onto a large white screen, the white screen rapidly appears yellow; the blue stimulus falling on the central retina causes inhibition of blue sensitivity in the periphery; hence, the white background will appear to be missing its blue light—white minus blue is a mixture of red and green—*i.e.,* yellow. Particularly interesting from this viewpoint are the phenomena of metacontrast; by this is meant the inductive effect of a primary light stimulus on the sensitivity of the eye to a previously presented light stimulus on an adjoining area of retina. It is a combination of temporal and spatial induction. The effect is produced by illuminating the two halves of a circular patch consecutively for a brief duration. If the left half only, for example, is illuminated for 10 milliseconds it produces a definite sensation of brightness. If, now, both halves are illuminated for the same period, but the right half from 20 to 50 milliseconds later, the left half of the field appears much darker than before and, near the centre, may be completely extinguished. The left field has thus been inhibited by the succeeding, nearby, stimulus. The right field, moreover, appears darker than when illuminated alone—it has been inhibited by the earlier stimulus (paracontrast).

Meta- and paracontrast

Flicker. Another visual phenomenon that brings out the importance of inhibition is the sensation evoked when a visual stimulus is repeated rapidly; for example, one may view a screen that is illuminated by a source of light the rays from which may be intercepted at regular intervals by rotating a sector of a circular screen in front of it. If the sector rotates slowly, a sensation of black followed by white is aroused; as the speed increases the sensation becomes one of flicker—*i.e.,* rapid fluctuations in brightness; finally, at a certain speed, called the critical fusion frequency, the sensation becomes continuous and the subject is unaware of the alterations in the illumination of the screen.

At high levels of luminance, when cone vision is employed, the fusion frequency is high, increasing with increasing luminance in a logarithmic fashion—the Ferry-Porter law—so that at high levels it may require 60 flashes per second to reach a continuous sensation. Under conditions of night, or scotopic, vision, the frequencies may be as low as four per second. The difference between rod and cone vision in this respect probably resides in the power of the eye to inhibit activity in cones rapidly, so that the sensation evoked by a single flash is cut off immediately, and this leaves the eye ready to respond to the next stimulus. By contrast, the response in the rod lasts so much longer that, when a new stimulus falls even a quarter of a second later, the difference in the state of the rods is insufficient to evoke a change in intensity of sensation; it merely prolongs it. One interesting feature of an intermittent stimulus is that the intensity of the sensation of brightness, when fusion is achieved, is dependent on the relative periods of light and darkness in the cycle, and this gives one a method of grading the effective luminance of a screen; one may keep the intensity of the illuminating source constant and merely vary the period of blackness in a cycle of black and white. The effective luminance will be the average luminance during a cycle; this is known as the Talbot-Plateau law.

Visual acuity. As has been stated, the ability to perceive detail is restricted in the dark-adapted retina when the illumination is such as to excite only the scotopic type

of vision; this is in spite of the high sensitivity of the retina to light under the same conditions. The power of distinguishing detail is essentially the power to resolve two stimuli separated in space, so that, if a grating of black lines on a white background is moved farther and farther away from an observer, a point is reached when he will be unable to distinguish this stimulus pattern from a uniformly gray sheet of paper. The angle subtended at the eye by the spacing between the lines at the point where they are just resolvable is called the resolving power of the eye; the reciprocal of this angle, in minutes of arc, is called the visual acuity. Thus, a visual acuity of unity indicates a power of resolving detail subtending one minute of arc at the eye; a visual acuity of two indicates a resolution of one-half minute, or 30 seconds of arc. The visual acuity depends strongly on the illumination of the test target, and this is true of both daylight (photopic) and night (scotopic) vision; thus, with a brightly illuminated target, with the surroundings equally brightly illuminated (the ideal condition), the visual acuity may be as high as two. When the illumination is reduced, the acuity falls so that, under ordinary conditions of daylight viewing, visual acuity is not much better than unity. Under scotopic conditions, the visual acuity may be only 0.04 so that lines would have to subtend about 25 minutes at the eye to be resolvable; this corresponds to a thickness of 4.4 centimetres (1.7 inches) at a distance of six metres (20 feet).

Resolving power and visual acuity

Measurement. In the laboratory, visual acuity is measured by the Landolt C, which is a circle with a break in it. The subject is asked to state where the break is when the figure is rotated to successive random positions. The size of the *C*, and thus of its break, is reduced until the subject makes more than an arbitrarily chosen percentage of mistakes. The angle subtended at the eye by the break in the *C* at this limit is taken as the resolving power of the eye. The testing of the eyes by the ophthalmologist or optometrist is essentially a determination of visual acuity; here the subject is presented with the Snellen chart, rows of letters whose details subtend progressively smaller angles at the eye. The row in which, say, five out of six letters are seen correctly is chosen as that which measures the visual acuity. If the details subtended one minute of arc, the visual acuity would be unity. The notation employed is somewhat obscure; a visual acuity of unity would be expressed as 6/6; an acuity of a half as 6/12, and so on; here the numerator is the viewing distance in metres from the chart and the denominator the distance at which details on the letters of the limiting row subtend one minute of arc at the eye.

The Snellen chart

Anatomical basis: the retinal mosaic. From an anatomical point of view one may expect the limit to resolving power to be imposed by the "grain" of the retinal mosaic in the same way that the size of the grains in a photographic emulsion imposes a limit to the accuracy with which detail may be photographed. Two white lines on a black ground, for example, could not be appreciated as distinct if their images fell on the same or adjacent sets of receptors. If a set of receptors intervened between the stimulated ones, there would be a basis for discrimination because the message sent to the central nervous system could be that two rows of receptors, separated by an unstimulated row, were sending messages to their bipolar cells. Thus, the limit to resolution, on this basis, should be the diameter of a foveal cone, or rather the angle subtended by this at the nodal point of the eye; this is about 30 seconds of arc and, in fact, corresponds with the best visual acuity attainable. If this grain of the retinal mosaic is to be the basis of resolution, however, one must postulate, in addition, a nervous mechanism that will transmit accurately the events taking place in the individual receptors, in this case the foveal cones; *i.e.*, there must be a one-to-one relationship between cones, bipolar cells, ganglion cells, and lateral geniculate cells so that what is called the local sign of the impulses from a given foveal cone may be obtained. It must be appreciated that restriction on convergence (or its reverse, spread) of messages may be achieved by inhibition; the anatomical connections may be there, but they may be made functionally inoperative by inhibition exerted by other neurons; thus, the horizontal and amacrine cells might well exert a restraining influence on certain junctions, thereby reducing the spread, or convergence, of messages, and it seems likely that the improvement in foveal visual acuity from one to two, brought about by increased luminance of the target and its surroundings, is achieved by an increase in inhibition that tends to make transmission one-to-one in the fovea.

It must be appreciated that true one-to-one connections in the retina do not exist; a cone, although making an exclusive type of synapse with a midget bipolar, may also make a less exclusive contact with a flat bipolar cell; furthermore, midget bipolars and cones are connected laterally by amacrine and horizontal cells so that it is most unlikely that a given optic nerve fibre carries messages from only a single cone. The one-to-one relationship may in fact exist under certain conditions, but that is because pathways from other receptors have been blocked or occluded by inhibitory processes that keep the line clear for a given cone.

The low visual acuity obtained in night, or rod, vision is now understandable. It has been pointed out that a high sensitivity to light is achieved by the convergence of rods on the higher neurons to allow spatial summation, and it is this convergence that interferes with the resolution of detail. If hundreds of rods converge on a single bipolar cell and if many bipolar cells converge on a single ganglion cell, it is understandable that the unit responsible for resolution may be very large and thus that the visual acuity is very small.

Scotopic acuity

The retinal image. It has been implied, in the comments on visual acuity, that the limiting factor is one of an anatomical arrangement of receptors and of their neural organization. A very important feature, however, must be the accuracy of the formation of an image of external objects by the optical system of the eye. It may be calculated, for example, that the image of a grating produces lines 0.5 micron wide on the retina, but this is on the basis of ideal geometrical optics; in fact, the optics of the eye are not perfect, while diffraction of light by its passage through the pupil further spoils the image. As a result of these defects, the image of a black and white grating on the retina is not sharp, the black lines being not completely black but gray because of spread of light from the white lines. (When the optical system of the eye is defective, moreover, as in nearsightedness, the imagery is worse, but this can be corrected by the use of appropriate lenses.) Physiologically, the eye effectively improves the retinal image by enhancing contrasts; thus, the image of a fine black line on a white background formed on the retina is not a sharply defined black line but a relatively wide band of varying degrees of grayness; yet to the observer the line appears sharply defined, and this is because of lateral inhibition, the receptors that receive most light tending to inhibit those that receive less; the result is a physiological "sharpening of the image," so that the eye often behaves as though the image were perfect. This applies to chromatic aberration, too, which should cause black and white objects to appear fringed with colour, yet, because of suppression of the chromatic responses, one is not aware of the coloured fringes that do in effect surround the images of objects in the external world.

The iris behaves as a diaphragm, modifying the amount of light entering the eye; probably of greater significance than control of the light entering the eye is the influence on aberrations of the optical system; the smaller the pupil the less serious, in general, are the aberrations. The smaller the pupil, however, the more serious become the effects of diffraction, so that a balance must be struck. Experimentally, it is found that at high luminances with pupils below three millimetres (0.12 inch) in diameter the visual acuity is not improved by further reduction of the diameter; increasing the pupil size beyond this reduces acuity, presumably because of the greater optical aberrations. It is interesting that when a subject is placed in a room that is darkened steadily, the size of the pupil increases, and the size attained for any given level of luminance is, in fact, optimal for visual acuity at this particular luminance. The reason that visual acuity increases with the larger pupils is that the extra light admitted into the eye compensates

The pupil

for the increased aberrations. When the gaze is fixed intently on an object for a long time, peripheral images that tend to disappear reappear immediately when the eyes are moved. This effect is called the Troxler phenomenon. To study it reproducibly it is necessary to use an optical device that ensures that the image of any object upon which the gaze is fixed will remain on the same part of the retina however the eyes move. Two investigators found, when they did this, that the stabilized retinal image tended to fade within a few seconds. It may be assumed that in normal vision the normal involuntary movements—the microsaccades and drifts mentioned earlier—keep the retinal image in sufficient movement to prevent the fading, which is essentially an example of sensory adaptation, the tendency for any receptive system to cease responding to a maintained stimulus.

Electrophysiology of the retina. *Neurological basis.* Subjective studies on human beings can traverse only a certain distance in the interpretation of visual phenomena; beyond this the standard electrophysiological techniques, which have been successful in unravelling the mechanisms of the central nervous system, must be applied to the eye; this, as repeatedly emphasized, is an outgrowth of the brain.

Receptive fields

Records from single optic nerve fibres of the frog and from the ganglion cell of the mammalian retina indicated three types of response. In the frog there were fibres that gave a discharge when a light was switched on the "on-fibres." Another group, the "off-fibres," remained inactive during illumination of the retina but gave a powerful discharge when the light was switched off. A third group, the "on-off fibres," gave discharges at "on" and "off" but were inactive during the period of illumination. The responses in the mammal were similar, but more complex than in the frog. The mammalian retina shows a background of activity in the dark, so that on- and off-effects are manifest as accentuations or diminutions of this normal discharge. In general, on-elements gave an increased discharge when the light was switched on, and an inhibition of the background discharge when the light was switched off. An off-element showed inhibition of the background discharge during illumination and a powerful discharge at off; this off-discharge is thus a release of inhibition and reveals unmistakably the inhibitory character of the response to illumination that takes place in some ganglion cells. Each ganglion cell or optic nerve fibre tested had a receptive field; and the area of frog's retina from which a single fibre could be activated varied with the intensity of the light stimulus. The largest field was obtained with the strongest stimulus, so that, in order that a light stimulus, falling at some distance away from the centre of the field, might affect this particular fibre it had to be much more intense than a light stimulus falling on the centre of the field. This means that some synaptic pathways are more favoured than others.

The mammalian receptive field is more complex, the more peripheral part of the field giving the opposite type of response to that given by the centre. Thus, if, at the centre of the field, the response was "on" (an on-centre field) the response to a stimulus farther away in the same fibre was at "off," and in an intermediate zone it was often mixed to give an on-off element. In order to characterize an element, therefore, it must be called on-centre or off-centre, with the meaning thereby that at the centre of its receptive field its response was at "on" or at "off," respectively, while in the periphery it was opposite. By studying the effects of small spot stimuli on centre and periphery separately and together, one investigator demonstrated a mutual inhibition between the two. A striking feature was the effect of adaptation; after dark adaptation the surrounding area of opposite activity became ineffective. In this sense, therefore, the receptive field shrinks, but, as it is a reduction in inhibitory activity between centre and periphery, it means, in fact, that the effective field can actually increase during dark adaptation—*i.e.*, the regions over which summation can occur—and this is exactly what is found in psychophysical experiments on dark adaptation.

Anatomical basis. The receptive field is essentially a measure of the number of receptors—rods or cones or a mixture of these—that make nervous connections with a single ganglion cell. The organization of centre and periphery implies that the receptors in the periphery of an on-centre cell tend to inhibit it, while those in the centre of the field tend to excite it, so that the effects of a uniform illumination covering the whole field tend to cancel out. This has an important physiological value, as it means, in effect, that the brain is not bombarded with an enormous number of unnecessary messages, as would be the case were every ganglion cell to send discharges along its optic nerve fibre as long as it was illuminated. Instead, the cell tends to respond to change—*i.e.*, the movement of a light or dark spot over the receptive field—and to give an especially prominent response, often when the spot passes from the periphery to the centre, or vice versa. Thus, the centre-periphery organization favours the detection of movement; in a similar way it favours the detection of contours because these give rise to differences in the illumination of the parts of the receptive fields. The anatomical basis of the arrangement presumably is given by the organization of the bipolar and amacrine cells in relation to the dendrites of the ganglion cell; it is interesting that the actual diameter of the centre of the receptive field of a ganglion cell is frequently equal to the area over which its dendrites spread; the periphery exerts its effects presumably by means of amacrine cells that are capable of connecting with bipolars over a wide area. These amacrine cells could exert an inhibitory action on the bipolar cells connected to the receptors of the central zone of the field, preventing them from responding to these receptors; in this case, the ganglion cell related to these bipolars would be of an on-centre and off-periphery type.

Inhibitory effect of periphery receptors

Direction-sensitive ganglion cells. When examining the receptive fields of rabbit ganglion cells, investigators found some that gave a maximal response when a moving spot of light passed in a certain "preferred" direction, while they gave no response at all when the spot passed in the opposite direction; in fact, the spontaneous activity of the cell was usually inhibited by this movement in the "null" direction. It may be assumed that the receptors connected with this type of ganglion cell are organized in a linear fashion, so that the stimulation of one receptor causes inhibition of a receptor adjacent to it. This inhibition would prevent the excitatory effect of light on the adjacent receptor from having a response when the movement was in the null direction, but would arrive too late at the adjacent receptor if the light was moving in the preferred direction.

The electroretinogram. If an electrode is placed on the cornea and another, indifferent electrode, placed, for example, in the mouth, illumination of the retina is followed by a succession of electrical changes; the record of these is the electroretinogram or ERG. Modern analysis has shown that the electrode on the cornea picks up changes in potential occurring successively at different levels of the retina, so that it is now possible to recognize, for example, the electrical changes occurring in the rods and cones—the receptor potentials—those occurring in the horizontal cells, and so on. In general, the electrical changes caused by the different types of cell tend to overlap in time, so that the record in the electroretinogram is only a faint and attenuated index to the actual changes; nevertheless, it has, in the past, been a most valuable tool for the analysis of retinal mechanisms. Thus, the most prominent wave—called the *b*-wave—is closely associated with discharge in the optic nerve, so that in animals, or man, the height of the *b*-wave can be used as an objective measure of the response to light. Hence, the sensitivity of the dark-adapted frog's retina to different wavelengths, as indicated by the heights of the *b*-waves, can be plotted against wavelength to give a typical scotopic sensitivity curve with a maximum at 5000 angstroms (one angstrom = 1×10^{-4} micron) corresponding to the maximum for absorption of rhodopsin.

Flicker. Electrophysiology has been used as a tool for the examination of the basic mechanism of flicker and fusion. The classical studies based on the electroretinogram indicated that the important feature that determines fusion in the cone-dominated retina is the inhibition of

the retina caused by each successive light flash, inhibition being indicated by the *a*-wave of the electroretinogram. In the rod-dominated retina—*e.g.*, in man under scotopic conditions— the *a*-wave is not prominent, and fusion depends simply on the tendency for the excitatory response to a flash to persist, the inhibitory effects of a succeeding stimulus being small. More modern methods of analysis, in which the discharges in single ganglion cells in response to repeated flashes are measured, have defined fairly precisely the nature of fusion, which, so far as the retinal message is concerned, is a condition in which the record from the ganglion cell becomes identical with the record observed in the ganglion cell during spontaneous discharge during constant illumination.

The nature of fusion

Visual acuity. Although the resolving power of the retina depends, in the last analysis, on the size and density of packing of the receptors in the retina, it is the neural organization of the receptors that determines whether the brain will be able to make use of this theoretical resolving power. It is therefore of interest to examine the responses of retinal ganglion cells to gratings, either projected as stationary images on to the receptive field or moved slowly across it. One group of investigators showed that ganglion cells of the cat differed in sensitivity to a given grating when the sensitivity was measured by the degree of contrast between the black and white lines of the grating necessary to evoke a measurable response in the ganglion cell. When the lines were made very fine (*i.e.*, the "grating-frequency" was high), a point was reached at which the ganglion cell failed to respond, however great the contrast; this measured the resolving power of the particular cell being investigated. The interesting feature of this work is that individual ganglion cells had a special sensitivity to particular grating-frequencies, as if the ganglion cells were "tuned" to particular frequencies, the frequencies being measured by the number of black and white lines in a given area of retina. When the same technique was applied to human subjects, the electrical changes recorded from the scalp being taken as a measure of the response, the same results were obtained.

Colour vision. The spectrum, obtained by refracting light through a prism, shows a number of characteristic regions of colour—red, orange, yellow, green, blue, indigo, and violet. These regions represent large numbers of individual wavelengths; thus, the red extends roughly from 7600 angstrom units to 6500; the yellow from 6300 to 5600; green from 5400 to 5000; blue from 5000 to 4200; and violet from 4200 to 4000. Thus, the limits of the visual spectrum are commonly given as 7600 to 4000 angstroms. In fact, however, the retina is sensitive to ultraviolet light to 3500 angstroms, the failure of the short wavelengths to stimulate vision being due to absorption by the ocular media. Again, if the infrared radiation is strong enough, wavelengths as long as 10,000–10,500 angstroms evoke a sensation of light.

Hue discriminations

Within the bands of the spectrum, subtle distinctions in hue may be appreciated. The power of the eye to discriminate light on the basis of its wavelength can be measured by projecting onto the two halves of a screen lights of different wavelengths. When the difference is very small—*e.g.*, five angstroms—no difference can be appreciated. As the difference is increased, a point is reached when the two halves of the screen appear differently coloured. The hue discrimination (hue is the quality of colour that is determined by wavelength) measured in this way varies with the region of the spectrum examined; thus, in the blue-green and yellow it is as low as 10 angstroms, but in the deep red and violet it may be 100 angstroms or more. Thus, the eye can discriminate several hundreds of different spectral bands, but the capacity is limited. If it is appreciated that there are a large number of nonspectral colours that may be made up by mixing the spectral wavelengths, and by diluting these with white light, the number of different colours that may be distinguished is high indeed.

Spectral sensitivity curve. At extremely low intensities of stimuli, when only rods are stimulated, the retina shows a variable sensitivity to light according to its wavelength, being most sensitive at about 5000 angstroms, the absorption maximum of the rod visual pigment, rhodopsin. In the light-adapted retina one may plot a similar type of curve, obtained by measuring the relative amounts of light energy of different wavelengths required to produce the same sensation of brightness; now the different stimuli appear coloured, but the subject is asked to ignore the colours and match them on the basis of their luminosity (brightness). This is carried out with a special instrument called the flicker-photometer. There is a characteristic shift in the maximum sensitivity from 5000 angstroms for scotopic (night) vision to 5550 angstroms for photopic (day) vision, the so-called Purkinje shift. It has been suggested that the cones have a pigment that shows a maximum of absorption at 5550 angstroms, but the phenomena of colour vision demand that there be three types of cone, with three separate pigments having maximum absorption in the red, green, and blue, so that it is more probable that the photopic luminosity curve is a reflection of the summated behaviour of the three types of cone rather than of one.

The Purkinje shift has an interesting psychophysical correlate; it may be observed, as evening draws on, that the luminosities of different colours of flowers in a garden change; the reds become much darker or black, while the blues become much brighter. What is happening is that, in this range of luminosities, called mesopic, both rods and cones are responding, and, as the rod responses become more pronounced—*i.e.*, as darkness increases—the rod luminosity scale prevails over that of the cones.

It may be assumed that the sensation of luminosity under any given condition is determined by certain ganglion cells that make connections to all three types of cone and also to rods; at extremely low levels of illumination their responses are determined by the activity aroused in the rods. As the luminance is increased, the ganglion cell is activated by both rods and cones, and so its luminosity curve is governed by both rod and cone activity. Finally, at extremely high luminances, when the rods are "saturated" and ceasing to respond, the luminosity curve is, in effect, compounded of the responses of all three types of cone.

Colour mixing. The fundamental principle of colour mixing was discovered by Isaac Newton when he found that white light separates spatially into its different component colours on passing through a prism. When the same light is passed through another prism, so that the individual bands of the spectrum are superimposed on each other, the sensation becomes one of white light. Thus, the retina, when white light falls on it, is really being exposed to all the wavelengths that make up the spectrum. Because these wavelengths fall simultaneously on the same receptors, the evoked sensation is one of white. If the wavelengths are spread out spatially, they evoke separate sensations, such as red or yellow, according to which receptors receive which bands of wavelengths. In fact, the sensation of white may be evoked by employing much fewer wavelengths than those in the spectrum: namely, by mixing three primary hues—red, green, and blue.

Three primary hues

Furthermore, any colour, be it a spectral hue or not, may be matched by a mixture of these three primaries, red, green, and blue, if their relative intensities are varied. Many of the colours of the spectrum can be matched by mixtures of only two of the primary colours, red and green; thus the sensations of red, orange, yellow, and green may be obtained by adding more and more green light to a red one.

To one accustomed to mixing pigments, and to mixing a blue pigment, for example, with yellow to obtain green, the statement that red plus green can give yellow or orange, or that blue plus yellow can give white, may sound strange. The mixing of pigments is essentially a subtractive process, however, as opposed to the additive process of throwing differently coloured lights on a white screen. Thus, a blue pigment is blue because it reflects mainly blue (and some green) light and absorbs red and yellow; and a yellow pigment reflects mainly yellow and some green and absorbs blue and red. When blue and yellow pigments are mixed, and white light falls on the mixture, all bands of colour are absorbed except for the green colour band.

Colour defectiveness. The colour-defective subject is

one whose wavelength discrimination apparatus is not as good as that of the majority of people, so that he sees many colours as identical that normal people would see as different. About one percent of males are dichromats; they can mix all the colours of the spectrum, as they see them, with only two primaries instead of three. Thus, the protanope (red blind) requires only blue and green to make his matches; since, for the normal (trichromatic) subject the various reds, oranges, yellows, and many greens are the result of mixing red and green, the protanope matches all these with a green. In other words, he is unable to distinguish all these hues from each other on the basis of their colour; if he distinguishes them, it is because of their different luminosity (brightness). The protanope matches white with a mixture of blue and green and is, in fact, unable to distinguish between white and bluish-green. The deuteranope (green blind) matches all colours with a mixture of red and blue; thus, his white is a mixture of red and blue that appears purple to a person with normal vision. The deuteranope also is unable to discriminate reds, oranges, yellows, and many greens, so that both types of dichromat are classed as red-green-blind. For the protanope, however, the spectrum is more limited because he is unable to appreciate red. The tritanope (blue blind) is rare, constituting only one in 13,000 to 65,000 of the population; because he is blue blind, his colour discrimination is best in the region of red to green, where that of the protanope and deuteranope is worse.

Responses of uniform population of receptors. The scotopic (night) visual system, mediated by rods, is unable to discriminate between different wavelengths; thus, a threshold stimulus of light with a wavelength of 4800 angstroms gives a sensation of light that is indistinguishable from that evoked by a wavelength of 5300 angstroms. If the intensities are increased, however, the lights evoke sensations of blue and green, respectively. Rods are unable to mediate wavelength, or colour, discrimination while the cones can because the rods form a homogeneous population, all containing the same photopigment, rhodopsin. Thus, the response of a nerve cell connected with a rod or group of rods will vary with the wavelength of light, and probably in the manner indicated by Figure 42, in which

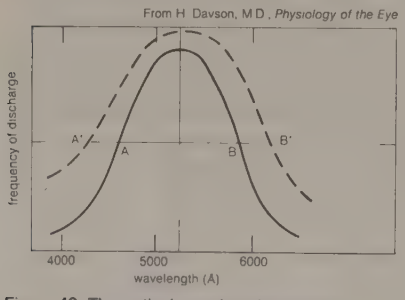


Figure 42: Theoretical wavelength response curve for a single receptor of *Limulus* retina. The maximum response will occur at 5200 Å, the absorptional maximum of the retinal pigment. With wavelengths of light corresponding to the points A and B, the responses will be identical, so that no discrimination between these two wavelengths is possible. When the intensity of the light is changed and a new curve is obtained (the dotted line), the responses at A' and B' are also identical.

the response, measured in frequency of discharge in the bipolar or ganglion cell, is plotted against the wavelength of the stimulating light. The curve is essentially similar to the absorption spectrum of rhodopsin when the same amount of energy is in each stimulus; thus, blue-green of 5000 angstroms has the most powerful effect because it is absorbed most efficiently, while violet and red have the smallest effects. In this sense, the rods behave as wavelength discriminators, but it is to be noted that there are pairs of wavelengths on each side of the peak to which the same response is obtained; thus, a blue of 4800 angstroms and a yellow of 6000 angstroms give the same discharge. Moreover, if the intensity of the stimulus is varied, a new curve is obtained (as in the dotted line of Figure 42), and now the same response is obtained with a high intensity of violet at 4000 angstroms as with blue at the lower intensity. In general, it is easy to show that, by varying the intensity of the stimulus of a single wavelength, all types of response may be obtained, so that the brain would never receive a message indicating, in a unique fashion, that the retina was stimulated with, say, green light of 5300 angstroms; the same message could be given by blue light of 4800 angstroms, red light of 6500 angstroms, and so on.

Why rods fail to discriminate between colours

Ideally, colour discrimination would require a large number of receptors specifically sensitive to small bands of the spectrum, but the number would have to be extremely large because the capacity for hue discrimination is extremely great, as has been indicated. In fact, however, the phenomena of colour mixing suggest that the number of receptors may be limited.

Young-Helmholtz theory. It was the phenomena of colour mixing that led Thomas Young in 1802 to postulate that there are three receptors, each one especially sensitive to one part of the spectrum; these receptors were thought to convey messages to the brain, and, depending on how strongly they were stimulated by the coloured light, the combined message would be interpreted as that due to the actual colour. The theory was developed by Hermann Ludwig Ferdinand von Helmholtz, and is called the Young-Helmholtz trichromatic theory. As expressed in modern terms, it is postulated that there are three types of cone in the retina, characterized by the presence of one of three different pigments, one absorbing preferentially in the red part of the spectrum, another in the green, and another in the blue. A coloured stimulus—*e.g.*, a yellow light—would stimulate the red and green receptors, but would have little effect on the blue; the combined sensation would be that of yellow, which would be matched by stimulating the eye with red and green lights in correct proportions of relative intensity. A given coloured stimulus would, in general, evoke responses in all three receptors, and it would be the pattern of these responses—*e.g.*, blue strongly, green less strongly, and red weakest—that would determine the quality of the sensation. The intensity of the sensation would be determined by the average frequencies of discharge in the receptors. Thus, increasing the intensity of the stimulus would clearly change the responses in all the receptors, but if they maintained the same pattern, the sensation of hue might remain unaltered and only that of intensity would change; the observer would say that the light was brighter but still bluish green. Thus, with several receptors, the possibility is reduced of confusion between stimuli of different intensity but the same wavelength composition; the system is not perfect because the laws of colour mixing show that the eye is incapable of certain types of discrimination, as, for example, between yellow and a mixture of red and green, but as a means of discriminating subtle changes in the environment the eye is a very satisfactory instrument.

The direct proof that the eye does contain three types of cone has been secured, but only relatively recently. This was done by examining the light emerging from the eye after reflection off the retina; in the dark-adapted eye the light emerging was deficient in blue light because this had been preferentially absorbed by the rhodopsin. In the light-adapted eye, when only cone pigments are absorbing light, the emerging light can be shown to be deficient in red and green light because of the absorption by pigments called erythrolabe and chlorolabe. Again, the light passing through individual cones of the excised human retina can be examined by a microscope device, and it was shown by such examination that cones were of three different kinds according to their preference for red, green, and blue lights.

The nervous messages. If the three types of cones respond differently to light stimuli, one may expect to find evidence for this difference in type of response by examining the electrophysiological changes taking place in the retina; ideally, one should like to place a microelectrode in or on a cone, then in or on its associated bipolar cell, and so on up the visual pathway. In the earliest studies, the optic nerve fibres of the frog were examined—*i.e.*, the axons of ganglion cells. The light-adapted retina was stimulated with wavelengths of light stretching across the

spectrum, and the responses in arbitrarily selected single fibres were examined. The responses to stimuli of the same energy but different wavelengths were plotted as frequency of discharge against wavelength, and the fibres fell into several categories, some giving what the investigator called a dominator response, the fibre responding to all wavelengths and giving a maximum response in the yellow-green at 5600 angstroms. Other fibres gave responses only over limited ranges of wavelengths, and their wavelengths of maximum response tended to be clustered in the red, green, and blue regions. The investigator called these modulators, and considered that the message in the dominator indicated to the brain the intensity of the stimulus—*i.e.*, it determined the sensation of brightness—while the modulators indicated the spectral composition of the stimulus, the combined messages in all the modulators resulting in a specific colour sensation. In the dark-adapted retina, when only rods were being stimulated, the response was of the dominator type, but this time the maximum response occurred with a wavelength of 5000 angstroms, the absorption maximum of rhodopsin.

Dominator impulse

A more careful examination of the responses in single fibres, especially in the fish, which has good colour vision, showed that things were not quite as simple as the original investigator had thought because, as has been seen, the response of a ganglion cell, when light falls on its receptive field in the retina, is not just a discharge of action potentials that ceases when the light is switched off. This type of response is rare; the most usual ganglion cell or optic nerve fibre has a receptive field organized in a concentric manner, so that a spot of light falling in the central part of the field produces a discharge, while a ring of light falling on the surrounding area has the opposite effect, giving an off-response—*i.e.*, giving a discharge only when the light is switched off. Such a ganglion cell would be called an on-centre-off-periphery unit; others behaved in the opposite way, being off-centre-on-periphery.

When these units are examined with coloured lights, and when care is taken to stimulate the centres and surrounding areas separately, an interesting feature emerges; the centre and surrounding areas usually have opposite or opponent responses. Thus, some may be found giving an on-response to red in the centre of the field and an off-response to green in the surrounding area, so that simultaneous stimulation of centre with red and surrounding area with green gives no response, the inhibitory effect of the off-type of response cancelling the excitatory effect of the on-type. With many other units the effects were more complex, the centre giving an on-response to red and an off-response to green, while the surrounding area gave an off-response to red and an on-response to green, and vice versa. This opponent organization probably subserves several functions. First, it enables the retina to emphasize differences of colour in adjacent parts of the field, especially when the boundary between them moves, as indeed it is continually doing in normal vision because of the small involuntary movements of the eyes. Second, it is useful in "keeping the retina quiet"; there are about one million optic nerve fibres, and if all these were discharging at once the problem of sorting out their messages, and making meaning of them, would be enormous; by this "opponence," diffuse white light falling on many of these chromatic units would have no effect because the inhibitory surrounding area cancelled the excitatory centre, or vice versa. When the light became coloured, however, the previously inactive units could come into activity.

These responses show that by the time the effect of light has passed out of the eye in the optic nerve the message is well colour-coded. Thus all the evidence points to the correctness of the Young-Helmholtz hypothesis with respect to the three-colour basis. The three types of receptor, responding to different regions of the spectrum in specific manners, transmit their effects to bipolar and horizontal cells. The latter neurons have been studied from the point of view of their colour-coding. The potentials recorded from them were called *S*-potentials; these were of two types, which classified them as responding to colour (*C*-units) and luminosity (*L*-units).

S-potentials

The *C*-type of cell gave an opponent type of response, in the sense that the electrical sign varied with the wavelength band, red and green having opponent effects on some cells, and blue and yellow on others. These responses reflect the connections of the horizontal cells to groups of different cones, the blue-yellow type, for example, having connections with blue and red and green cones, while the red-green would have connections only with red and green cones.

Lateral geniculate cells. As indicated above, the cells at the next stage, the ganglion cells, give a fairly precisely coded set of messages indicating the chromatic (colour) quality and the luminosity (brightness) of the stimulus, organized in such a way, however, as to facilitate the discrimination of contrast. At higher stages—*e.g.*, in the cells of the lateral geniculate body—this emphasis on opponence, or contrast, is maintained and extended; thus, several types of cell have been described that differ in accordance with the organization of their receptive fields from the colour aspect; some were very similar to ganglion cells, while others differed in certain respects. Some showed no opponence between colours when centre and periphery were compared, so that if a red light on the periphery caused inhibition, green and blue light would also do so. Others had no centre-periphery organization, the receptive field consisting of only a central spot; different colours had different effects on this spot; and so on.

In the cerebral cortex there is the same type of opponence with many units, but because cortical cells require stimuli of definite shape and often are not activated by simple spot stimuli, early studies carried out before these requirements were known probably failed to elucidate the true chromatic requirements of these high-order neurons. In general, the responses are what might be predicted on the basis of connections made to lateral geniculate neurons having the chromatic responses already known. Thus the final awareness of colour probably depends on the bombardment of certain higher-order cortical neurons by groups of primary cortical neurons, each group sending a different message by virtue of the connections it makes to groups of cones, connections mediated, of course, through the neurons of the retina and lateral geniculate body.

The photochemical process. For the energy of light to exert its effect it must be absorbed; it has been stated above that the action-spectrum for vision (the sensitivity of the eye to light) in the completely dark-adapted eye has a maximum in the region of 5000 angstroms, and that this corresponds with the maximum of absorption of light by the pigment, rhodopsin, extracted from the dark-adapted retina of the same species. The chemical nature of rhodopsin must now be examined, as well as its localization in the rod and the changes it undergoes in response to the absorption of light. It must be appreciated at the outset that the amount of light energy absorbed by a single rod at the threshold for vision is extremely small—namely, one quantum—and this is quite insufficient to provide the energy required to cause an electrical change in the membrane of the rod that will be propagated from the point of absorption of the light to the rod spherule (which takes part in the synapse between rod and bipolar cell). There must, therefore, be a chemical amplification process taking place within the rod, and the absorption of a quantum must be viewed as the trigger that sets off other changes, which in turn provide the required amount of energy.

Rhodopsin. Visual purple, or rhodopsin, is a chromoprotein, a protein, opsin, with an attached chromatophore ("pigment-bearing") molecule that gives it its colour—*i.e.*, that allows it to absorb light in the visible part of the spectrum. In the absence of such a chromatophore, the protein would only absorb in the ultraviolet and so would appear colourless to the eye. The chromatophore group was identified as retinal, which is the substance formed by oxidation of vitamin A; on prolonged exposure of the eye to light, retinal can be found, free from the protein opsin, in the retina. When the eye is allowed to remain in the dark, the rhodopsin is regenerated by the joining up of retinal with opsin. Thus one may write:

Opsin and retinal

$$\text{rhodopsin} \rightleftharpoons \text{retinal} + \text{opsin}.$$

The incidence of light on the retina causes the reaction

to go to the right (that is, causes rhodopsin to form retinal plus opsin), and this photochemical change causes the sensation of light. The process is reversed by a thermal—*i.e.*, non-photochemical—reaction, so that for any given light intensity a steady state is reached with the regenerative process just keeping pace with the photochemical bleaching. Dark adaptation, or one element in it, is the regenerative process. The change in the rhodopsin molecule that leads to its bleaching—*i.e.*, the splitting off of the retinal molecule—takes place in a succession of steps; and there is reason to believe that the electrical change in the rod that eventually evokes the sensation of light occurs at a stage well before the splitting off of the retinal. One may describe as a transduction process the chemical events that take place between the absorption of light and the electrical event, whatever that may be; the rod behaves as a transducer in that it converts light into electrical or neural energy.

Prelumirhodopsin

The transduction process. Immediately after absorption of a quantum, the rhodopsin molecule is changed into a substance called prelumirhodopsin, recognized by its different colour from that of rhodopsin; this product is so highly unstable that at body temperature it is converted, without further absorption of light, into a series of products. These changes may be arrested by cooling the solution to −195° C (−319° F), at which temperature prelumirhodopsin remains stable; on warming to −140° C (−220° F) prelumirhodopsin becomes lumirhodopsin, with a slightly different colour; on warming further, successive changes are permitted until finally retinal is split off from the opsin to give a yellow solution. The important point to appreciate is that only at this stage is the chromatophore group split off; the earlier products have involved some change in the structure of the chromoprotein, but not so extreme as to break off the retinal. The precise nature of these changes is not yet completely elucidated, but the most fundamental one—namely, that occurring immediately after absorption of the quantum—has been shown to consist in a change in shape of the retinal molecule while it is still attached to opsin.

Thus retinal, like vitamin A, can exist in several forms because of the double bonds in its carbon chain—the socalled *cis-trans* isomerism. In other words, the same group of atoms constituting the retinal molecule can be twisted into a number of different shapes, although the sequence of the atoms is unaltered. While attached to the opsin molecule in the form of rhodopsin, the retinal has a shape called 11-*cis*, being somewhat folded, while on conversion to prelumirhodopsin the retinal has a straighter shape called all-*trans;* the process is called one of photoisomerization, the absorption of light energy causing the molecule to twist into a new shape. Having suffered this alteration in shape, the retinal presumably causes some instability in the opsin, making it, too, change its shape, and thereby exposing to the medium in which it is bathed chemical groupings that were previously shielded by being enveloped in the centre of the molecule. It may be assumed that these changes in shape induce alterations in the light-absorbing character of the molecule that permit the recognition of the new forms of molecule represented by lumirhodopsin, metarhodopsins I and II, and so on.

Photoisomerization

The final change is more drastic because it involves the complete splitting off of the retinal; an earlier stage—namely, the conversion of metarhodopsin I to metarhodopsin II—has been shown recently to involve a bodily change in position of the retinal, which in rhodopsin is linked to the lipid (fatty) portion of the molecule, whereas in metarhodopsin II it is found to have become attached to an amino acid in the backbone-chain of the protein (amino acids are subunits of proteins). Thus, in its native unilluminated state, retinal is attached to a lipid, which is presumably linked to the protein, so that rhodopsin is more properly called a chromolipoprotein rather than a chromoprotein. The outer segments of the rods are, as has been stated, constituted by membranous disks, and it is well established that the material from which these membranes are constructed is predominantly lipid, so that one may envisage the rhodopsin molecules as being, in fact, part of the membrane structure. The techniques used for extraction presumably tear the molecules from the main body of the lipid, but some of the lipid remains with the protein and retinal to constitute the link holding these two parts together.

Within the retina these chemical changes are all reversible, so that when a steady light is maintained on the retina the latter will contain a mixture of several or all of the intermediate compounds. In the dark, all will be gradually reconverted to rhodopsin. Because lack of vitamin A, from which retinal is derived, causes night blindness, some of the retinal must get lost from the eye to the general circulation; and it is actually replaced by the cells of the pigment epithelium, which are closely associated with the rods.

As to which of these chemical changes acts as the trigger for vision, there is some doubt. The discovery that the transition from metarhodopsin I to metarhodopsin II involves an actual shift of the retinal part of the molecule from linkage to lipid to linkage to protein reinforces the belief that this particular shift is sufficient to lead ultimately to electrical discharges in the optic nerve.

Cone pigments. So far as colour vision is concerned, the changes that take place in the three cone pigments have not been analyzed, simply because, so far, they have defied isolation, presumably because their concentrations are so much less than that of the rod pigment.

THE HIGHER VISUAL CENTRES

The visual pathway. The axons of the ganglion cells converge on the region of the retina called the papilla or optic disk. They leave the globe as the optic nerve, in which they maintain an orderly arrangement in the sense that fibres from the macular zone of the retina occupy the central portion, the fibres from the temporal half of the retina take up a concentric position, and so on; when outside the orbit, there is a partial decussation (crossover). The fibres from the nasal halves of each retina cross to the opposite side of the brain, while those from the temporal halves remain uncrossed. This partial decussation is called the chiasma. The optic nerves after this point are called

Partial decussation

By permission from Eugene Wolff, *Anatomy of the Eye and Orbit*, London, H K Lewis & Co Ltd

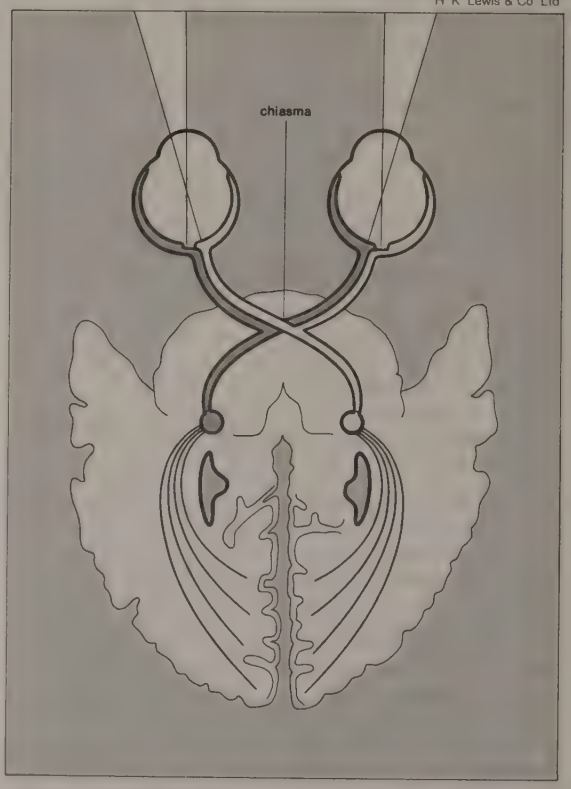


Figure 43: *Visual pathways.*
Fibres from the nasal side of the retina cross over in the chiasma to join the uncrossed fibres of the temporal half of the retina.

the optic tracts, containing nerve fibres from both retinas. The result of the partial decussation is that an object in, say, the right-hand visual field produces effects in the two eyes that are transmitted to the left-hand side of the brain only. With cutaneous (skin) sensation there is a complete crossing-over of the sensory pathway; thus, information from the right half of the body, and the right visual field, is all conveyed to the left-hand part of the brain by the time that it has reached the diencephalon (the posterior part of the forebrain, Figure 43).

Fusion of retinal images. Partial decussation is an arrangement that serves the needs of frontally directed eyes and permits binocular vision, which consists in the fusion of the responses of both eyes to a single object—more loosely, one speaks of the fusion of the retinal images. In many lower mammals, with laterally directed eyes and therefore limited binocular vision, the degree of decussation is much greater, so that in the rat, for example, practically all of the optic nerve fibres pass to the opposite side of the brain.

The fibres of the optic tracts relay their messages to nerve cells in those parts of the diencephalon called the lateral geniculate bodies, and from the lateral geniculate bodies the messages are relayed to nerve cells in the occipital cortex of the same side. (The occipital cortex is the outer substance in the posterior portion of the brain.)

The visual field. If one eye is fixed on a point in space, the visual field for this eye may be thought of as the part of a surface of a sphere on to which all visible objects are projected. The limits to this field will be determined by the sensitivity and extent of the retina and the accessibility of light rays from the environment. Experimentally or clinically, the field is measured on a perimeter, a device for ascertaining the point on a given meridian where a white spot just appears or disappears from vision when moved along this meridian. (A meridian is a curve on the surface of a sphere that is formed by the intersection of the sphere surface and a plane passing through the centre of the sphere.) The field is recorded on a chart, illustrated by Figure 44. On the nasal side, the field is restricted to about 60° from the midline. This is due to the obstruction caused by the nose, since the retina extends nearly as far forward on the temporal side of the globe as on the nasal side. It is customary to refer to the binocular visual field as that common to the two eyes, the uniocular field being the extreme temporal (outside) region peculiar to each eye. It will be clear from the field of the single eye shown in Figure 44 that the binocular field is determined in the horizontal meridian by the nasal field of each eye, and so will amount to about 60° to either side of the vertical meridian.

From H. Davson and M.G. Eggleton (eds.), *Principles of Human Physiology*

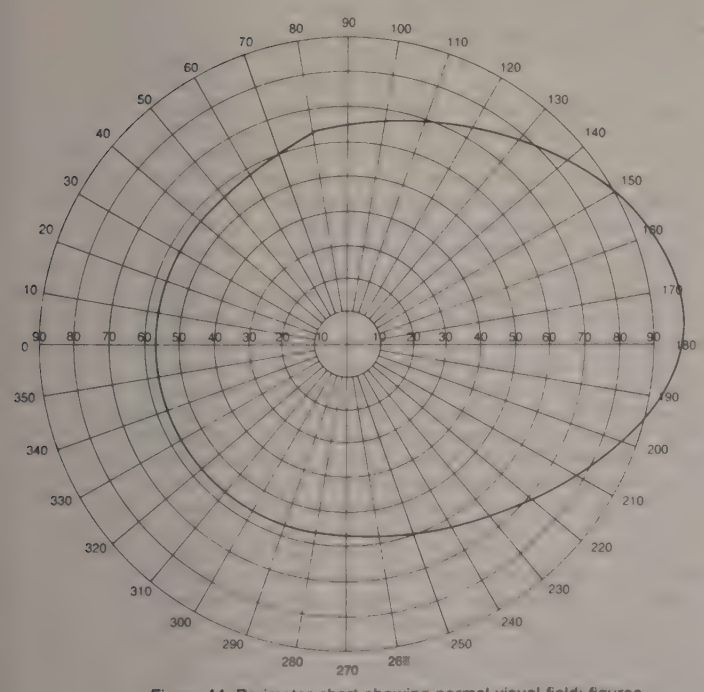


Figure 44: Perimeter chart showing normal visual field; figures on the perimeter indicate degrees of arc.

Lateral geniculate body. The dorsal (posterior) nucleus of the lateral geniculate body, where the optic tract fibres relay, has six layers, and the crossed fibres relay in layers 1, 4, and 6, while the uncrossed relay in layers 2, 3, and 5; thus, at this level, the impulses from the two eyes are kept separate, and when the discharges in geniculate neurons are recorded electrically it is rare to find any responding to stimuli in both eyes.

Striate area. The optic tract fibres make synapses with nerve cells in the respective layers of the lateral geniculate body, and the axons of these third-order nerve cells pass upward to the calcarine fissure (a furrow) in each occipital lobe of the cerebral cortex. This area is called the striate area because of bands of white fibres—axons from nerve cells in the retina—that run through it. It is also identified as Brodmann's area 17. It is at this level that the impulses from the separate eyes meet at common cortical neurons, or nerve cells, so that when the discharges in single cortical neurons are recorded it is usual to find that they respond to light falling in one or the other eye. It is probable that it is when the retinal messages have reached this level of the central nervous system, and not before, that the human subject becomes aware of the visual stimulus, since destruction of the area causes absolute blindness in man. Because of the partial decussation, however, the removal of only one striate cortex will not cause complete blindness in either eye, since only messages from two halves of the retinas will have been blocked; the same will be true if one optic tract is severed or one lateral geniculate body is destroyed. The result of such lesions will be half-blindness, or hemianopia, the messages from one half of the visual field being obliterated.

Point of awareness of visual stimulus

Pupillary pathways. Some of the fibres in the optic tracts do not relay in the lateral geniculate bodies but pass instead to a midbrain region—the pretectal centre—where they mediate (transmit) reflex alterations in the size of the pupil. Thus, in bright light, the pupils are constricted; this happens by virtue of the pupillary light reflex mediated by these special nerve fibres. Removal of the occipital cortex, although it causes blindness in the opposite visual field, does not destroy the reaction of the pupils to light; if the optic nerve is cut, however, the eye will be both completely blind and also unreactive to light falling on this eye. The pupil of the blind eye will react to light falling on the other eye by virtue of a decussation in the pupillary reflex pathway.

Point-to-point representation. Because of the ordered manner in which the optic tract fibres relay in the lateral geniculate bodies and from there pass in an orderly fashion to the striate area, when a given point on the retina is stimulated, the response recorded electrically in either the lateral geniculate body or the striate area is localized to a small region characteristic for that particular retinal spot. When the whole retinal field is stimulated in this point-to-point way, and the positions on the geniculate or striate gray matter on which the responses occur are plotted, it is possible to plot on these regions of the brain maps of the retinal fields or, more usually, maps of the visual fields.

Visuopsychic or circumstriate areas. Area 17, the striate area, is the primary visual centre in the sense that, in primates at any rate, all of the geniculate fibres project onto it and none projects onto another region of the cortex. There are two other areas containing neurons that have close connections with the eye; these are the parastriate and peristriate areas, or Brodmann's areas 18 and 19, respectively, in close anatomical relationship to one another and to area 17. They are secondary visual areas in the sense that messages are relayed from area 17 to area 18 and from area 18 to area 19, and, because area 17 does not relay to regions beyond area 18, these circumstriate areas are the means whereby visual information is brought into relation with more remote parts of the cortex. Thus in writing, the eyes direct the activities of the fingers, which are controlled

Peristriate and parastriate areas

by a region of the frontal cortex, so that one may presume that visual information is relayed to this frontal region. In the monkey, bilateral destruction of the areas causes irrecoverable loss of a learned visual discrimination, but this can be relearned after the operation. In man, lesions in this region are said to cause disturbances in spatial orientation and stereoscopic vision, but much more knowledge is required before specific functions can be attributed to these circumstriate areas, if, indeed, this is possible.

Integration of the retinal halves. The two halves of the retina, and thus of the visual field, are represented on opposite cerebral hemispheres, but the visual field is perceived as a unity and hence one would expect an intimate connection between the two visual cortical areas.

Corpus callosum. The great bulk of the connections between the two sides of the cerebral mantle are made by the interhemispheric commissure (the point of union between the two hemispheres of the cerebrum) called the corpus callosum, which is made up of neurons and their axons and dendrites that make synapses with cortical neurons on symmetrically related points of the hemispheres. Thus, electrical stimulation of a point on one hemisphere usually gives rise to a response on a symmetrically related point on the other, by virtue of these callosal connections. The striate area is an exception, however, and it is by virtue of the connections of the striate neurons with the area 18 neurons that this integration occurs, the two areas 18 on opposite hemispheres being linked by the corpus callosum.

Stereopsis in the midline. Usually stereopsis, or perception of depth, is possible by the use of a single hemisphere because the images of the same object formed by right and left eyes are projected to the same hemisphere; however, if the gaze is fixed on a distant point and a pin is placed in line with this but closer to the observer, a stereoscopic perception of the distant point and the pin can be achieved by the fusion of disparate images of the pin, but the images of the pin actually fall on opposite retinal halves, so that this fusion must be brought about by way of the corpus callosum.

Callosal transfer. In experimental animals it is possible, by section of the chiasma, to ensure that visual impulses from one eye pass only to one hemisphere. If this is done, an animal trained to respond to a given pattern and permitted to use only one eye during the training is just as efficient, when fully trained, in making the discrimination with the other eye. There has thus been a callosal transfer of the learning so that the hemisphere that was not directly involved in the learning process can react as well as that directly involved. If the corpus callosum is also sectioned, this transfer is impossible, so that the animal, trained with one eye, must be trained again if it is to carry out the task with the other eye only.

Effect of cutting the corpus callosum

Superior colliculi. The visual pathway so far described is called the geniculostriate pathway, and in man it may well be the exclusive one from a functional aspect because lesions in this pathway lead to blindness. Nevertheless, many of the optic tract fibres, even in man, relay in the superior colliculi, a paired formation on the roof of the midbrain. From the colliculi there is no relay to the cortex, so that any responses brought about by this pathway do not involve the cortex. In man, as has been said, lesions in the striate area, which would of course leave the collicular centres intact, cause blindness, so that the visual fibres in these centres serve no obvious function. In lower animals, including primates, removal of the striate areas does not cause complete blindness; in fact, it is often difficult to determine any visual impairment from a study of the behaviour of the animals. Thus, in reptiles and birds, vision is barely affected, so that a pigeon that has been subjected to the operation can fly and avoid obstacles as well as a normal one. In rodents, such as the rabbit, removal of the occipital lobes causes some impairment of vision, but the animal can perform such feats as avoiding obstacles when running and recognizing food by sight. In the monkey, the effects are more serious, but the animal can be trained to discriminate lights of different intensity and even the shapes of objects, provided that these are kept in continual motion. It seems likely, then, that it is the visual pathway through the colliculi that permits the use of the eyes in the absence of visual cortex, although the connections of the optic tract fibres with the pulvinar of the thalamus (an area in the diencephalon), established in some animals, may well permit the use of regions of the cortex other than those denoted as visual.

SOME PERCEPTUAL ASPECTS OF VISION

So far, the visual process has been considered from rather elementary aspects; the ability to detect light and changes in its intensity, and to discriminate colour and form. It is now time to deal with more complex features, particularly some phenomena of binocular vision. It will then be in order to return to the electrophysiology of the visual pathway to see how some of the phenomena can be interpreted.

Projection of the retina. Objects are perceived in definite positions in space—positions definite in relation to each other and to the percipient. The first problem is to analyze the physiological basis for this spatial perception or, as it is expressed, the projection of the retina into space.

Relative positions of objects. The perception of the positions of objects in relation to each other is essentially a geometrical problem. Take, for the present, the perception of these relationships by one eye, monocular perception: a group of objects, as in Figure 45, produces images on the

From H. Davson, M.D., *Physiology of the Eye*

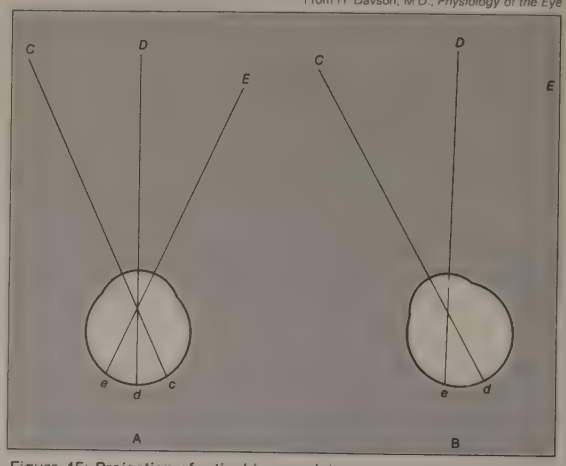


Figure 45: *Projection of retinal images into space.* (A) A group of objects, *C,D,E,* produces images on the retina, *c,d,e;* the retinal images are projected outward in space toward the points evoking them. (B) The eye has moved to the left so that the image of *c* falls on *d,* previously projected to *D* but now projected to *C; E* no longer produces an image on the retina.

retina in a certain fixed geometrical relationship; for the perception of the fact that *C* is to the left of *D,* that *D* is to the left of *E,* and so on, it is necessary that the incidence of images at *c, d,* and *e* on the retina be interpreted in a similar, but, of course, inverted geometrical relationship. The neural requirements for this interpretation are (1) that the retina be built up of elements that behave as units throughout their conducting system to the visual cortex, and (2) that the retinal elements have "local signs." The local sign could represent an innate disposition or could result from experience—the association of the direction of objects in space, as determined by such evidence as that provided by touch, with the retinal pattern of stimulation. In neurophysiological terms, the retinal elements are said to be connected to cortical cells, each being specific for a given element, so that when a given cortical cell is excited the awareness is of a specific local sign. Studies of the projection of the retina on the cerebral cortex have confirmed this.

The retinal stimuli at *c, d,* and *e* in Figure 45 are appreciated as objects outside the eye, the retina is said to be projected into space, and the field of vision is thus the projection of the retina through the nodal point. (In Figure 45 the nodal point is the point of intersection of *Cc, Dd,* and *Ee.*) It will be seen that the geometrical relationship between objects and retinal stimuli is reversed; in the retina *c* is to the right of *d,* and so on.

Position in relation to observer. The recognition of the directions of objects in relation to the observer is more complex. If the eye (Figure 45B) is turned to the left, the image of *C* falls on the retinal point *d*, so that if *d* were always projected into the same direction in space, *C* would appear to be in *D*'s place. In practice, one knows that *C* is perceived as fixed in space in spite of the movements of the eye; hence, the direction of projection of a retinal point is constantly modified to take into account movements of the eye; this may be called psychological compensation. It will be seen that correct projection is achieved by projecting the stimulated retinal point through the nodal point of the eye. Movements of the eye caused by movements of the head must be similarly compensated. As a result, any point in space remains fixed in spite of movements of the eye and head. Given this system of compensated projection, the recognition of direction in relation to the individual is now feasible. *D* may be said to be due north or, more vaguely, "over there"; when the head is turned, since *D* is perceived to be in the same place, it is still due north or "over there." In some circumstances, the human subject makes an error in projecting his retinal image, so that the object giving rise to the image appears to be in a different place from its true one; the image is said to be falsely projected. If the eye is moved passively, for example, by pulling on the conjunctiva with forceps, the subject has the impression that objects in the outside world are moving in a direction opposite to that of the eye.

False projection

The apparent movement of an afterimage, when the eye moves, is an excellent illustration of psychological compensation. A retinal stimulus, being normally projected through the nodal point, is projected into different points in space as the eye moves; an afterimage can be considered to be the manifestation of a continued retinal impulse, and its projection changes as the eye moves. The afterimage thus appears to move in the same direction as that of the movement of the eye. Whether the drift of an afterimage across the field of view is entirely due to eye movements is difficult to say. One certainly has the impression that the eye is chasing the afterimage.

Visual estimates. *The directions of lines.* So far, consideration has been given to the problem of estimating the positions of points in relation to each other and to the percipient. The estimate of the directions of lines involves no really new principles, since, if two points, *A* and *B*, are exactly localized, the direction of the line *AB* can be appreciated. As will be seen, the organization of the neural connections of the retina and higher visual pathway is such as to favour the accurate recognition of direction; for the moment, the question of the maintenance of a frame of reference must be considered, in the sense that a map has vertical and horizontal lines with which to compare other directions. In fact, the vertical and horizontal meridians of the retina seem to be specialized as frames of reference; the accuracy with which a human subject can estimate whether a line is vertical or horizontal is very great.

An important point in this connection is that of the effects of eye movements on interpretation of the directions of lines because, when the eye moves to positions different from the primary straight-ahead position, the images of vertical lines will not necessarily fall on its vertical meridian. This can be due to an actual torsion of the eye about its anteroposterior (fore and aft) axis or to distortion of the retinal image. This means, then, that the line on the retina that corresponds to verticality in one position of the eye does not correspond to verticality in another, so that, once again, the space representation centre must take account not only of the retinal elements that have been stimulated but also of the corollary motor discharge.

Effects of eye movements

Comparison of lengths. The influence of the movements of the eyes in the estimation of length was emphasized by Helmholtz. An accurate comparison of the lengths of two parallel lines *AB* and *CD* can be made, whereas if an attempt is made to compare the nonparallel lines *A'B'* and *C'D'*, quite large errors occur. According to Helmholtz, the eye fixates first the point *A*, and the line *AB* falls along a definite row of receptors, thereby indicating its length. The eye is now moved to fixate *C*, and if the image of *CD* falls along the same set of receptors the length of *CD* is said to be the same as that of *AB*. Such a movement of the eye is not feasible with lines that are not parallel. Similarly, the parallelism, or otherwise, of pairs of lines can be perceived accurately because on moving the eye over the lines the distance between them must remain the same.

Fairly accurate estimates of relative size may be made, nevertheless, without movements of the eyes. If two equal lines are observed simultaneously, the one with direct fixation and the other with peripheral vision, their images fall, of course, on different parts of the retina; if the images were equally long it could be stated that a certain length of stimulated retina was interpreted as a certain length of line in space. It is probable that this is roughly the basis on which rapid estimates of length depend, although there are such complications as the fact that the retina is curved so that lines of equal length in different parts of the retina do not produce images of equal length on the retina.

Optical illusions. Many instances have been cited of well-defined and consistent errors in visual estimates under special conditions. There is probably no single factor by which the errors can be explained, but the tendency for distinctly perceptible differences to appear larger than those more vaguely perceived is important.

The perception of depth. *Monocular cues.* The image of the external world on the retina is essentially flat or two-dimensional, and yet it is possible to appreciate its three-dimensional character with remarkable precision; to a great extent this is by virtue of the simultaneous presentation of different aspects of the world to the two eyes, but even when the subject views the world with a single eye it does not appear flat to him and he can, in fact, make reasonable estimates of the relative positions of objects in all three dimensions. Examples of monocular cues are the apparent movements of objects in relation to each other when the head is moved. Objects nearer the observer move in relation to more distant points in the opposite direction to the movement of the head. Perspective, by which is meant the changed appearance of an object when it is viewed from different angles, is another important clue to depth. Thus the projected retinal image of an object in space may be represented as a series of lines on a plane—*e.g.*, a box—these lines, however, are not a unique representation of the box because the same lines could be used to convey the impression of a perfectly flat object with the lines drawn on it, or of a rectangular, but not cubical, box viewed at a different angle. In order that a three-dimensional object be correctly represented to the subject on a two-dimensional surface, he must know what the object is; *i.e.*, it must be familiar to him. Thus a bicycle is a familiar object. If it is viewed at an angle from the observer the wheels seem elliptical and apparently differ in size. Because the observer knows that the wheels are circular and of the same size, he perceives depth in a two-dimensional pattern of lines. The perception of depth in a two-dimensional pattern thus depends greatly on experience—the knowledge of the true shape of things when viewed in a certain way. Other cues are light and shade, overlapping of contours, and relative sizes of familiar objects.

Perspective

Binocular vision. The cues to depth mentioned above are essentially uniocular; they would permit the appreciation of three-dimensional space with a single eye. When two eyes are employed, two additional factors play a role, the one not very important—namely, the act of convergence or divergence of the eyes—and the other very important—namely, the stereoscopic perception of depth by virtue of the dissimilarity of the images presented by a three-dimensional object, or array of objects, to the separate eyes.

When a three-dimensional object or array is examined binocularly, the nearer points or objects require greater convergence for fixation than the more distant points or objects, so that this provides a cue to the three-dimensional character of the presentation. It is by no means a necessary cue, since presentation of the array for such a short time that movements of the eyes cannot occur still permits the three-dimensional perception, which is achieved under these conditions by virtue of the dissimilar images received by the two retinas.

A stereogram contains two drawings of a three-dimensional object taken from different angles, chosen such that the pictures are right- and left-eyed views of the object. When the stereogram is placed in a stereoscope, an optical device for enabling the two separate pictures to be fused and seen single, the impression created is one of a three-dimensional object. The perception is immediate, and is not a matter of interpretation. Clearly, with the stereoscope the situation is simulated as it normally occurs. To appreciate the full implications of the stereoscopic perceptual process, one must examine some simpler aspects of binocular vision.

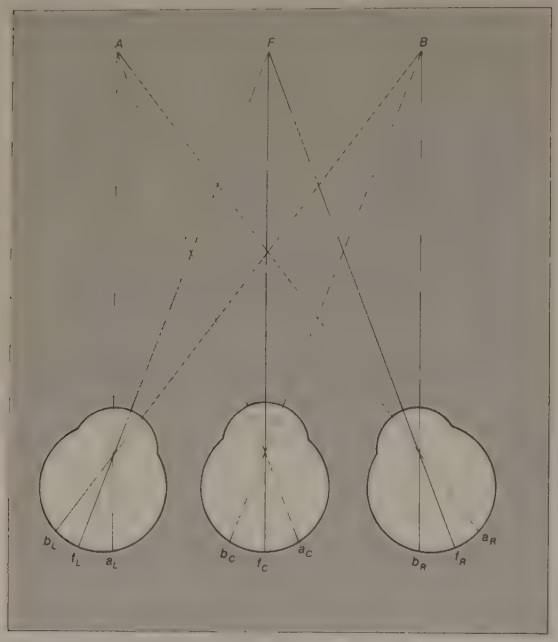


Figure 46: *The Cyclopean system of projection.* The images of the points *F*, *A*, and *B* on the two retinas are transposed to the retina of a hypothetical eye midway between the two. The pairs of images, a_L and a_R, b_L and b_R, and so on, coincide on the cyclopean retina indicating that the stimulated retina points are projected to a common direction (see text).

Figure 46 illustrates the situation in which a subject is fixating (fixing his gaze on) the point F so that the images of F fall on the foveal (retinal) points f_L and f_R, respectively. F is seen as a single point because the retinal points f_L and f_R are projected to the same point in space, and the projection is such that the subject says that the point F is straight in front of him, although it is to the right of his left eye and to the left of his right eye. The two eyes in this case are behaving as a single eye, "the cyclopean eye," situated in the centre of the forehead, and one may represent the projection of the two separate retinal points, f_L and f_R, as the single projection of the point f_C of the cyclopean eye. As will be seen, the cyclopean eye is a useful concept in consideration of certain aspects of stereoscopic vision.

Corresponding points

The points f_L and f_R may be defined as corresponding points because they have the same retinal direction values. The images formed by the points A and B, in the same frontal plane as F, fall on a_L and a_R and b_L and b_R; once again the pairs of retinal points are projected to the same points, namely, to A and B, and they are treated as being on the left and right of F, respectively. On the cyclopean projection, they may be said to be localized by the outward projections of a_C and b_C, respectively.

In Figure 47, the subject is once again fixing the point F, but the point A is now no longer in the same frontal plane as the point F, but closer to the observer. The images of F fall on corresponding points and are projected to a single point in front. The images of A, on a_L and a_R, do not fall on corresponding points and are, in fact, projected into space in different directions, as indicated by the cyclopean projection. This means that A is seen simultaneously at two different places, a phenomenon called physiological diplopia, and this in fact does happen, as can be seen by fixing one's gaze on a distant point and holding a pencil fairly close to the face; with a little practice the two images of the pencil can be distinguished. Thus, when the eyes are directed into the distance the objects closer to the observer are seen double, although one of the double images of any pair is usually suppressed. To return to Figures 46 and 47, F and A in Figure 46 are seen single and in the same plane because their images each fall on corresponding points. F is seen single and A double in Figure 47 because the images of A fall on noncorresponding, or disparate, points. A is appreciated as being closer to the observer than F in Figure 47 by virtue of these double images but, in general, although it is retinal disparity that creates the percept of three-dimensional space, it is not necessarily the formation of double images, since if the disparity is not large the point will be seen single, and this single point will appear to be in a different frontal plane from that containing the fixation point.

To appreciate the nature of this stereoscopic perception one must examine what is meant by corresponding points in a little more detail. In general, it seems that the two retinas are, indeed, organized in such a way that pairs of points are projected innately to the same point in space, and the horopter is defined as the outward projection of these pairs. One may represent this approximately by a sphere passing through the fixation point, or, if one confines attention to the fixation plane, it may be represented by the so-called Vieth-Müller horopter circle, as illustrated in Figure 48. On this basis, the corresponding points are arranged with strict symmetry, and each pair projects to a single point in space on the horopter circle. Theoretically, then, all points on the circle passing through the fixation point, F, will be seen single, and the point X will be seen double because it will be projected by the left eye to F and by the right eye to A. The actual situation is somewhat more complex than this, since experimentally the horopter turns out to have different shapes according to how close the fixation point is to the observer. The point to appreciate, however, is that the experimentally determined line, be it circular or straight or elliptical, is such that when points are placed on it they all appear to be in the same frontal plane—*i.e.*, there is no stereoscopic perception of depth when one views these points—and one may say that this is because the images of points on the horopter fall on corresponding points of the two retinas.

In Figure 49, to the left, are two eyes viewing an arrow

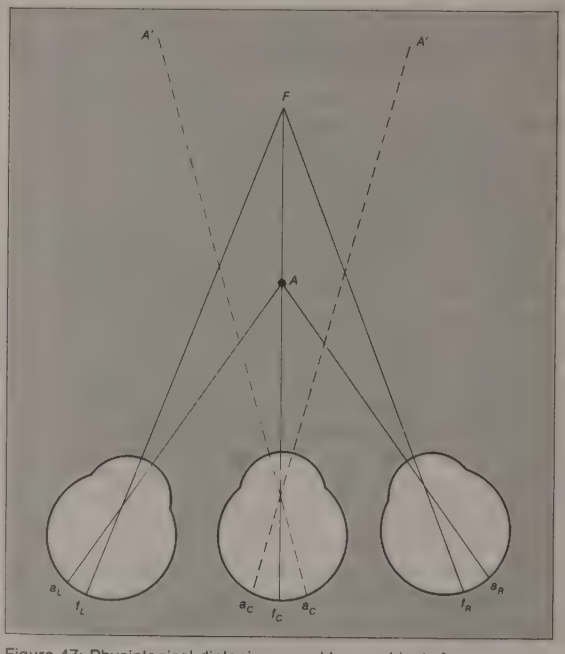


Figure 47: Physiological diplopia caused by an object, *A*, closer to the observer than the fixation point, *F*. The images of *A* fall on disparate or noncorresponding points on the two retinas, and these are projected to different points *A'* and *A'* (see text).

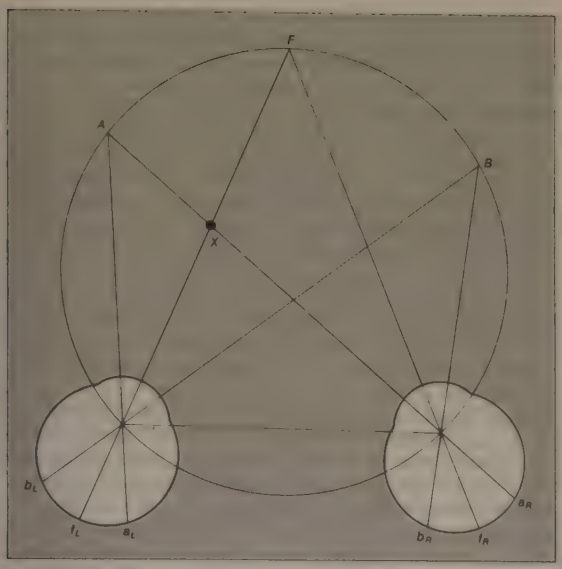


Figure 48: *The Vieth-Müller horopter circle.*
F is the fixation point. If corresponding points are symmetrically distributed about the foveas, the points in space in the fixation plane, whose images fall on the corresponding points, lie on the circle. The images of the point *X* lie on disparate points (see text).
From H. Davson, M.D., *Physiology of the Eye*

Fusion of disparate images

lying in the frontal plane—*i.e.*, with no stereopsis—and to the right the arrow is inclined into the third dimension—*i.e.*, it tends to point toward the observer. All points on the arrow are, in fact, seen single under both conditions, and yet it is clear from the right-hand figure that, if the gaze is fixed on *A*, the images of *B′* will fall on noncorresponding points. *B′* is not seen double but, instead, the noncorresponding points, b'_L and b'_R, are projected to a common point *B′* and a stereoscopic percept is achieved. Thus the noncorresponding, or disparate, points on the retinas can be projected to a single point, and it is essentially this fusion of disparate images by the brain that creates the impression of depth. If the point *B′* were brought much closer to the eyes, its images would fall on such disparate points that fusion would no longer be possible, and *B′* would be seen double, or one double image would be suppressed. There is thus a certain zone of disparity that, if not exceeded, allows fusion of disparate points. This is called Panum's fusional area; it is the area on one retina such that any point in it will fuse with a single point on the other retina.

To return to the stereoscopic perception of three-dimensional space, one may recapitulate that it is because the two eyes receive different images of the same object that the stereoscopic percept happens; when the two images of the object are identical, then, except under very special conditions, the object has no three-dimensionality. A special condition is given by a uniformly illuminated sphere; this is three-dimensional, but the observer would

From H. Davson, M.D., *Physiology of the Eye*

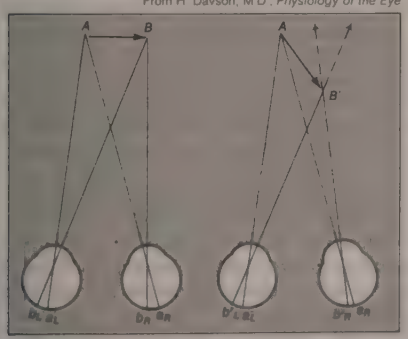


Figure 49: The distinction between corresponding points (a_L and a_R, b_L and b_R) and points that do not correspond (b'_L and b'_R; see text).

have to use special cues to discriminate this from a flat disk lying in the frontal plane. Such a cue might be the different degree of convergence of the eyes required to fixate the centre from that required to fixate the periphery, or the different degree of accommodation.

The difference in the two aspects of the same object (or group of objects), measured as the instantaneous parallax, is illustrated in Figure 50. *B* is closer to the observer than *A*; the fact is perceived stereoscopically because the line *AB* subtends different angles at the two eyes, and the instantaneous parallax is measured by the difference between the angles *a* and *b*. The binocular parallax of any point in space is given by the angle subtended at it by the line joining the nodal points of the two eyes; hence, the binocular parallax of *A* is *a*; that of *B* is *b*; the instantaneous parallax is thus the difference of binocular parallax of the two points considered.

Accuracy of stereoscopic perception

If one places three vertical wires in front of an observer in the frontal plane, one may move the middle one in front of, or behind, the plane containing the other two and ask the subject to say when he perceives that it is out of the plane; under correct experimental conditions the only cue will be the difference of binocular parallax, and it is found that the minimum difference is remarkably small, of the order of five seconds of arc, corresponding to a disparity of retinal images far smaller than the diameter of a single cone. With two editions of the same book, it is not possible, by mere inspection, to detect that a given line of print was not printed from the same type as the same line in the other book. If the two lines in question are placed in the stereoscope, it is found that some letters

From Hugh Davson, M.D. *Physiology of the Eye*

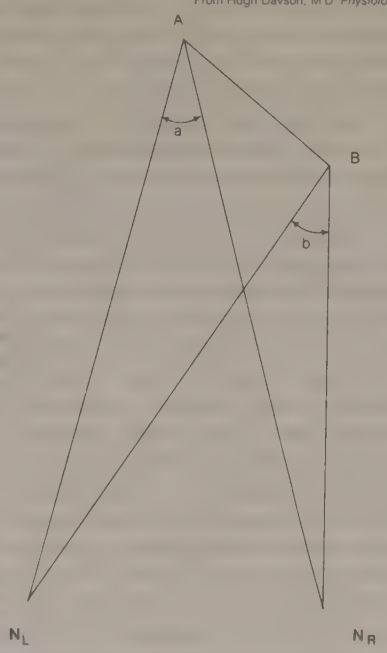


Figure 50: *Binocular and instantaneous parallax.*
N_L and N_R are the nodal points of the left and right eyes, respectively (see text).

appear to float in space, a stereoscopic impression created by the minute differences in size, shape, and relative position of the letters in the two lines. The stereoscope may thus be used to detect whether a bank note has been forged, whether two coins have been stamped by the same die, and so on.

The stereoscopic appearance obtained by regarding two differently coloured, but otherwise identical, plane pictures with the two eyes separately, is probably due to chromatic differences of magnification. If the left eye, for example, views a plane picture through a red glass and the right eye views the same picture through a blue glass, an illusion of solidity results. Chromatic difference in magnification causes the images on the two retinas to be slightly different in size, so that the images of any point on the picture

do not fall on corresponding points; the conditions for a stereoscopic illusion are thus present.

Retinal rivalry. Stereoscopic perception results from the presentation to the two eyes of different images of the same object; if two pictures that cannot possibly be related as two aspects of the same three-dimensional object are presented to the two eyes, single vision may, under some conditions, be obtained, but the phenomenon of retinal rivalry enters. Thus, if the letter *F* occupies one side of a stereogram, and *L* the other, the two letters can be fused by the eyes to give the letter *E;* the letters *F* and *L* cannot, however, by any stretch of the imagination be regarded as left and right aspects of a real object in space, so that the final percept is not three-dimensional, and, moreover, it is not a unitary percept in the sense used in this discussion; great difficulty is experienced in retaining the appearance of the letter *E*, the two separate images, *F* and *L*, tending to float apart. This is a mode of binocular vision that may be more appropriately called simultaneous perception; the two images are seen simultaneously, and it is by superimposition, rather than fusion, that the illusion of the letter *E* is created. More frequent than superimposition is the situation in which one or the other image is completely suppressed; thus, if the right eye views a vertical black bar and the left eye a horizontal one, the binocular percept is not that of a cross; usually the subject is aware of the vertical bar alone or the horizontal bar alone. Moreover, there is a fairly characteristic rhythm of suppression, or alternation of dominance, as it is called.

Ocular dominance. Retinal rivalry may be viewed as the competition of the retinal fields for attention; such a notion leads to the concept of ocular dominance—the condition when one retinal image habitually compels attention at the expense of the other. While there seems little doubt that a person may use one eye in preference to the other in acts requiring monocular vision—*e.g.,* in aiming a rifle—it seems doubtful whether, in the normal individual, ocular dominance is really an important factor in the final awareness of the two retinal images. Where the retinal images overlap, stereoscopic perception is possible and the two fields, in this region, are combined into a single three-dimensional percept. In the extreme temporal fields (*i.e.,* at the outside of the fields of vision), entirely different objects are seen by the two eyes, and the selection of what is to dominate the awareness at any moment depends largely on the interest it arouses; as a result, the complete field of view is filled in and one is not aware of what objects are seen by only one eye. Where the fields overlap, and different objects are seen by the two eyes—*e.g.,* on looking through a window the bars may obscure some objects as seen by one eye but not as seen by the other—the final percept is determined by the need to make something intelligible out of the combined fields. Thus, the left eye may see a chimney pot on a house, while the other eye sees the bar of a window in its place; the final perceptual pattern involves the simultaneous awareness of both the bar and the chimney pot because the retinal images have meaning only if both are present in consciousness. So long as the individual retinal images can be regarded as the visual tokens of an actual arrangement of objects, it is possible to obtain a single percept, and there seems no reason to suppose that the final percept will be greatly influenced by the dominance of one or other eye. When a single percept is impossible, retinal rivalry enters; this is essentially an alternation of awareness of the two fields—the subject apparently makes attempts to find something intelligible in the combined presentation by suppressing first one field and then the other—and certainly it would be incorrect to speak of ocular dominance as an absolute and invariable imposition of a single field on awareness, since this does not occur. Dominance, however, has a well-defined physiological meaning in so far as certain cells of the cerebral cortex may be activated exclusively by one eye, either because the other eye makes no neural connections with it or because the influence of the other eye is dominant.

The choice of what is perceived

Ocular dominance

Binocular brightness sensation. When the two eyes are presented with differently illuminated objects or surfaces some interesting phenomena emerge. Thus fusion may give rise to a sensation of lustre. In other instances, rivalry takes place, the one or other picture being suppressed, while in still others the brightness sensation is intermediate between those of the two pictures. This gives rise to the paradox whereby a monocularly viewed white surface appears brighter than when it is viewed binocularly in such a way that one eye views it directly and the other through a dark glass. In this second case the eyes are receiving more light, but because the sensation is determined by both eyes, the result is one that would be obtained were one eye to look at a less luminous surface.

ELECTROPHYSIOLOGY OF THE VISUAL CENTRES

To elucidate the functions of the various stages in the visual pathway, one must examine the responses to a retinal light-stimulus of the individual neurons at the different stages.

Ganglion cells. The main features of the responses of ganglion cells have already been touched upon. These have receptive fields that indicate a dual type of connection with the rods and cones, as indicated by the centre-periphery organization. A spot of light falling on receptors in the centre of this field may provoke a discharge in the ganglion cell or its optic nerve fibre; it is called an on-response and consists usually in an increase in the background discharge occurring in darkness. If a spot of light falls on a ring of retina surrounding this central region, the effect is one of inhibition of the background while the light is on, and as soon as it is switched off there is a pronounced discharge, the off-response. Other ganglion cells have been shown to have a directional sensitivity, responding to a moving spot of light only if this moves in a preferred direction and showing inhibition of background discharge if movement is in the null direction.

Geniculate neurons. In general, the lateral geniculate neuron is characterized by an accentuation of the centre-periphery arrangement, so that the two parts of the receptive field tend to cancel each other out completely when stimulated together, by contrast with the ganglion cell in which one or another would predominate. Thus, when the retina is illuminated uniformly there is little response in the geniculate cells because of this cancellation. This represents a useful elaboration of the messages from the retina because, to the animal, uniformity is uninteresting; it is the nonuniformity created by a contour or a moving object that is of interest, and the brain is therefore spared from being bombarded by unnecessary information that would result if every receptor response were transmitted to the brain.

Cortical neurons. When investigators made records of responses from neurons in area 17 there was an interesting change in the nature of the receptive fields; there was still the organization into excitatory (on) and inhibitory (off) zones, but these were linearly arranged, so that the best stimulus for evoking a response was a line, either white on black or black on white. When this line fell on the retina in a definite direction, and on a definite part of the retina, there was, say, an on-response, while if it fell on adjacent areas there was an off-response. Changing the orientation of the line by as little as 15° could completely abolish the responses. The simplest interpretation of this type of receptive field is based on the connection of the cortical cell with a set of geniculate cells with their receptive fields arranged linearly.

Eye dominance. Most of these units (*i.e.,* cortical cells plus connections) could be excited by a light stimulus falling on either eye, although there was usually dominance of one eye, in the sense that its response was greater; when both eyes were stimulated together, the effects summated. In general, then, when a large number of units are studied, a certain proportion are fired by one eye alone, others by the opposite eye alone, others by both eyes with dominance of one or other eye, while still others respond only when both eyes are stimulated. It is interesting that when kittens are deprived of the use of one eye from birth for several months, this deprived eye is virtually blind and the distribution of dominance in the cortical neurons is changed dramatically; if the left eye is deprived, the right hemispherical cortical neurons show a marked fall

Variations in response

in dominance by the left eye, and an increase by the right eye. Thus, the ability of the eye to make use of cortical neurons is not fully developed at birth.

Cortical architecture. When an electrode is directed downward into the cortex it picks up responses in individual units at successive depths; units having the same directional sensitivity are arranged in columns so that the receptive fields of successive neurons are similarly oriented. When units were classified on the basis of eye dominance, a similar vertical distribution of units was found, overlapping with those based on directional preference. The columns for eye preference were about one millimetre wide, but those for directional preference were considerably finer. This columnar organization of cortical cells is not peculiar to the visual area.

Complex neurons. The cortical units (cells) described above, with receptive fields organized on a linear basis, have been called simple units in contrast to complex and hypercomplex units. Four types of complex units have been described; as with the simple units, the orientation of a slit stimulus (that is, a line) is of the utmost importance for obtaining maximal response, but unlike the situation with the simple unit, the position on the retina is unimportant. This type of unit makes abstractions of a higher order, responding to direction of orientation but not to position. It is this type of neuron that would be concerned, for example, with determining the verticality or horizontality of lines in space. Space does not permit of a description of the receptive field of a hypercomplex cell, but in general its features could be explained on the basis of connections with complex cells.

Stereoscopic vision. Of special interest is the behaviour of binocularly driven (stimulated) cortical cells, since their responses provide a clue to the fusion of retinal images. The cortical nerve cell receiving impulses emanating from both retinas must select those parts of the two retinal images that are the images of the same point on an object; second, for stereopsis, the nerve cell must assess the small displacements from exact symmetry that give the binocular parallax. In experiments, maximal response was often obtained only when the stimuli fell on disparate parts of the two retinas; these cortical cells were obviously disparity detectors, in contrast to others that gave maximal response when the stimuli fell on strictly symmetrically related parts of the two retinas—*i.e.,* on corresponding points. When successive units, during penetration of the electrode, were recorded, it was found that those requiring the same degree of disparity for maximal response were arranged in columns, as with direction sensitivity, so that, in effect, all these nerve cells were responding to stimuli in a strip of space at a definite distance from the fixation point. (H.Da.)

Eye diseases and visual disorders

The human visual apparatus includes the eyeball or globe, and its socket, or orbit, and auxiliary structures such as the lids and the muscles that control eye movement. These organs and their normal functioning are covered in detail above. This section briefly describes the more common diseases of the eye and its associated structures, and the methods used in examination and diagnosis; it also indicates the treatment and prognosis. The first part deals with conditions affecting the orbit, lids, and external eye, and the second with diseases of structures within the globe. Later sections deal with injuries, ocular conditions associated with general disease, disorders of vision, methods of examining the eye, and devices for correcting visual defects.

THE OUTER EYE AND AUXILIARY STRUCTURES

The orbit and lacrimal apparatus. The orbit is the bony cavity in the skull that contains the globe of the eye, the muscles that move the eye, the lacrimal gland, and the blood vessels and nerves required to supply these structures. The remaining space within the orbit is filled with a fatty pad that acts as a cushion for the eye and allows free movement of the globe. In old age this pad of fat tends to atrophy so that the globe recedes, causing the sunken appearance often seen in old people.

Inflammatory conditions of the orbit. As the bone that separates the orbit from the nose and the nasal sinuses is rather thin, infection sometimes spreads from the nasal sinuses into the orbit, causing the orbital tissue to swell and the eye to protrude. The condition is serious because of the possibility that the infection may spread into the cranial cavity along the pathways of the cranial nerves that enter the orbit to reach the eye. Infections can also spread to the cranial cavity by way of the blood vessels that lie in the upper part of the orbit, from lesions such as a boil on the skin of the lids or face in the neighbourhood of the eye. Large doses of an antibiotic such as penicillin, given immediately, in most cases eliminate such infections. The lacrimal glands, the small glands that secrete tears and are behind the outer part of each upper lid, are rarely inflamed but may become so as a complication of mumps. Inflammations of the lacrimal sac are much more common. The lacrimal or tear sac lies in a hollow at the corner of the eye in the front part of the nasal wall of the orbit; under normal conditions tears run along the edges of the lids toward the nose and are drained through two tiny holes connected by small tubes to the upper part of the lacrimal sac. The lower part of the sac is connected to the nose by a duct, the nasolacrimal duct, and infection may ascend this passage from the nose and cause an acute painful swelling at the inner corner of the eye. Blockage of the nasolacrimal duct prevents the passage of tears into the nose and results in a watering eye. Such a blockage, which is nearly always accompanied by chronic inflammation in the lacrimal sac, is usually best treated by an operation in which a new opening from the lacrimal sac to the nasal cavity is made.

Infections of nasal origin

Tumours of the orbit. Tumours in this area are comparatively rare, the most common being a tumour of the lacrimal gland. If the tumour is behind the globe in the optic nerve, it will cause a slow and gradual protrusion of the eye; such an abnormal position of the eye may prevent ocular movements from being coordinated with those of the normal eye, and the images of the two eyes, which are normally fused, may separate and give rise to double vision (or diplopia).

The lids. *Inflammatory conditions.* The chronic inflammation of the lid margins known as blepharitis is a common and distressing condition. The inflammation may be mild and consist simply of redness of the lid margin with scaling of the skin, or more severe, affecting the follicles of the eyelashes and leading to their destruction and distortion. Both types tend to be associated with greasiness of the skin and dandruff. The skin of the lids is particularly sensitive to allergic processes, and itching and scaling of the lids is a common reaction to drugs or cosmetics applied to the eye of a sensitized person.

Another common inflammatory condition of the lid is a sty; *i.e.,* an infection of a lash follicle—the sheath of the eyelash root—corresponding to a boil on the skin elsewhere. It starts as a painful swelling of the whole lid, so that at first it may be difficult to find a localized lesion; but soon one area becomes more swollen and, as pus forms, a yellow point associated with an eyelash can be seen near the lid margin. A rather similar appearance can be produced by an inflammation of the meibomian glands; *i.e.,* tiny glands in the thickness of the lid, opening on the lid margin. As the glands are embedded in tough fibrous tissue, the pain and reaction may be more severe than in an ordinary sty. Examination of the internal surface of the lid will show a red velvety area with a central yellow spot through which pus will later discharge. Sometimes the meibomian glands suffer from a chronic infection, and a painless firm lump appears in the lid and slowly increases in size. The skin can be moved freely over the surface of the lump, showing that the latter is in the deeper tissue of the lid. The inner surface of the lid will show a grayish area surrounded by a little inflammation. The lesion is treated by making an incision on the inner surface of the lid, and scraping out the contents.

Herpes zoster (shingles) may affect the skin of the eyelids and is of particular importance because the cornea (the transparent covering of the front of the eyeball) and the inner eye may also be affected. The condition starts with

pain and redness of one part of the forehead and the eyelids of the same side. Vesicles, or small blisters, form later in the affected area. The pain may be severe, and some constitutional disturbance is usual.

Ectropion and entropion

Displacements of the lid. Malpositioning of the lid is common in elderly people, and although not serious in itself gives rise to considerable discomfort and irritation. The commonest condition is called ectropion, in which the lower lid falls away from the globe so that the tears overflow the lid. This constant wetting of the skin of the lower lid excoriates the skin and causes it to retract, which in turn increases the tendency of the lid to turn out. In the early stages, massage of the lid and the use of a bland ointment on the skin may help, but usually some plastic procedure is necessary to bring the lid back into its normal position.

The opposite condition is entropion, in which the lid turns inward and the lashes cause much irritation by rubbing on the eye. Unlike ectropion, it may affect either the upper or the lower lid. It may be caused by scarring of the deeper tissues of the lid following infection, or may be due to senile changes in muscle tone in which a band of fibres of the circular muscle surrounding the lids contracts more strongly than the peripheral fibres, thus tending to turn the lid inward. Surgical treatment is required to restore the lid to its normal position.

Tumours. Benign overgrowths of the blood vessels, called hemangiomas, may occur in the lids and give rise to soft bluish swellings that can be reduced by pressure over them. They are present at birth and tend to grow rapidly in the first few years of life. Often they disappear spontaneously, but they can be treated by surgical removal or the insertion of radioactive material. Simple overgrowths of skin, called papillomas, are common along the lid margin but require no special treatment except excision for cosmetic reasons.

The lids and the skin of the nose near the inner margins of the lids are common sites for the development of skin cancer in older people; the most usual type, called a rodent ulcer, starts as a small nodule in the skin that gradually enlarges and breaks down to form an ulcer with a hard base and rolled edges. Bleeding may occur from the base of the ulcer. Although rodent ulcers are malignant in the sense that they destroy tissue locally, they do not spread to distant areas of the body by means of the lymph system or the blood vessels. A more serious malignant tumour of the lids is a carcinoma, which develops as a more irregular ulcer of the lid and may spread to involve the underlying bone. If left untreated the growth may spread to the lymph nodes in front of the ear or under the jaw.

Squint. In the lower vertebrates such as the fish the eyes are situated on either side of the head, to give the maximum view of the surroundings and an early warning of the presence of predators. The field of vision of each eye is separate except for a narrow sector immediately in front of the animal, where the two visual fields overlap. From the evolutionary point of view the improved judgment of distance obtained by viewing an object with both eyes conferred considerable biological advantage in the struggle for survival. In the higher animals, particularly the predatory species of birds and mammals, binocular vision became more and more important and structural changes in the placement of the eyes in the head permitted a larger overlap of the two visual fields, until the situation was reached in the higher mammals in which the visual axes—that is, the line of direct sight—became parallel. This desirable visual function has been fully attained in man. The structural changes necessary to bring this about have, however, lagged behind the function, and the geometrical axes of most eyes are still slightly divergent (*i.e.*, the two eyes at rest are directed slightly away from the nose). The bony structure of the orbit has lagged even further behind, so that the axes of the two orbits make an angle of about 45°.

Development of binocular vision; definition of squint

It is in fact the function of using two eyes together that keeps the optic axes straight in a normal person. If, for example, one eye becomes blind, it tends to revert to an anatomical position of rest in line with the axis of the orbit. A blind eye will therefore appear to be diverging. The visual axes can remain straight only if each eye has reasonably good vision, the ocular muscles can move the eyes in the required direction of gaze, and the complex neuromuscular reflexes required to coordinate the movements of the two eyes are intact. Failure to maintain the visual axes parallel may therefore result from a visual defect in one or both eyes, a muscular defect resulting in loss of normal movement of the eye, or a defect in the central nervous system involving the coordinating nervous pathways. A true squint is a condition in which the visual axes are no longer parallel. An apparent squint may be seen in children as the result of prominent skin folds in the inner sides of the eyes, which make the eyes appear to be converging (*i.e.*, appear to cross). These skin folds usually disappear when the bony structure of the nose develops more fully.

Clinically, squints are divided into concomitant, in which the abnormal angle between the visual axes remains constant in all positions of gaze, and paralytic, in which the angle of squint varies with the direction of gaze. The commonest type of squint is the convergent concomitant type seen in small children (*i.e.*, the children are consistently or intermittently cross-eyed). It is usually first noticed between the ages of one and two and may be precipitated by a systemic disease such as measles. There is often a family history of squint. Children of this age are particularly interested in objects close to them, and in order to view an object clearly at close range two things are necessary: first, the visual axes must converge, so that both eyes can view the same object; and second, the focus of the eye must be adjusted for near vision. The link between convergence of the eyes and focussing, or accommodation, is very strong, and normally the two actions work in harmony. Most small children, however, are long-sighted, which means that, in order to see clearly close to, they have to exert an extra amount of accommodative effort. As accommodation and convergence are closely linked, the extra effort of accommodation tends to produce an overconvergence; but, provided that the visual acuity of each eye is normal and the motor control of the eyes is normal, this tendency is controlled. If the vision of one eye is reduced, for example by disease or an error of refraction, binocular vision breaks down and overconvergence occurs.

Once parallelism of the visual axes is lost, the image of objects no longer lies on a familiar area of retina, and instead of the images from the two eyes being fused into one, two images are perceived. This condition of double vision, or diplopia, is intolerable to the child, who reacts by "suppressing" the image from the squinting eye. If the suppression is allowed to continue, the central vision of the affected eye drops rapidly to a low level, so that even if the original disturbance that started the squint is corrected, this loss of vision, or amblyopia, of the squinting eye will prevent the restoration of normal binocular vision and thus perpetuate the squint. The longer the suppression is allowed to continue, the less likely is the child to regain normal vision in the squinting eye. Covering the good eye will usually encourage the recovery of the suppressed vision but must be started as soon as the squint is noticed. Any refractive error present—any defect that prevents light rays from focussing properly on the retina—must be corrected by glasses, and retraining of the binocular reflexes can be aided by special exercises. Early treatment on these lines may be all that is necessary, but if the visual axes are still abnormal surgery of the extraocular muscles will be required to correct the deviation.

Double vision; suppression of one image

Paralysis of one of the muscles that control the movement of the eyes results in limitation of movement of the globe in the direction of action of the muscle, with the result that double vision with separation of the images occurs on attempts to move the eye in this direction.

As earlier stated, accommodation and convergence are normally perfectly linked together so that the movements of the two eyes bring the visual axes to the point of focus. In many people this balance is not quite perfect and the eyes tend to converge or diverge too much for a given distance—a condition known as heterophoria.

The conjunctiva. The marine origin of the human species is betrayed by the need for the anterior surface

of the eye to be bathed in salt water. A thin membrane lines the lids and covers the anterior surface of the globe, forming a sac, the conjunctival sac, the contents of which are lubricated by the tear glands. This warm, moist habitat provides a suitable environment for the growth of bacteria and other organisms, leading to conjunctivitis, inflammation of the conjunctiva. Bacterial conjunctivitis starts with a feeling of grittiness and discomfort, the eye is red and there is a discharge from it. The discharge is particularly noticeable after sleep, when the lids may be stuck together by the exudate on the lashes. Vision is not affected except by the strands of mucus, which can be blinked away from the cornea. Antibiotic drops usually clear the condition in a few days. Vernal conjunctivitis or spring catarrh is, as its name suggests, an allergic condition occurring in the early summer; it is more common in young people and probably results from sensitivity to external irritants such as dust and pollen. It usually responds well to treatment with drops of corticosteroid hormone.

Chronic conjunctivitis, causing a gritty feeling, with redness of the eyes and a slight mucoid discharge, is a common condition, the cause of which may be difficult to find. Often there is an infective element such as a chronic inflammation of the lid margin, and sometimes the condition is allergic and may result from sensitivity to cosmetics or to drugs applied to the eye. An unsuspected foreign body or a deficiency of tear secretion may cause similar symptoms.

Viral conjunctivitis. With the enormous increase in the use of antibiotics since the 1940s, bacterial infections in general are becoming less common. This is also true of infections of the eye, and in most western countries bacterial conjuctivitis is now less common than viral infection. Viruses tend to attack the cornea as well as the conjunctiva; the infection is contagious and may be responsible for outbreaks of epidemic keratoconjunctivitis (inflammation of the cornea and the conjunctiva). The onset is acute, with redness and swelling of the eye and lids and a tender swelling of the lymph node in front of the ear.

Trachoma

Trachoma can truly be described as one of the scourges of mankind. Although rare in England and North America, it is the largest single cause of blindness in the world as a whole. Widespread in some Middle East countries, it has become more common in Asia, India, Central and South America, and Africa. It occurs sporadically in southern and eastern Europe. The agent responsible has now been isolated and shown to belong to the group of organisms known as chlamydiae. They occupy a taxonomic position between bacteria and true viruses and unlike the latter are susceptible to treatment with sulfonamides and some antibiotics. The disease is contagious and thrives where populations are crowded together in poor hygienic surroundings. Shortage of water for washing, and the myriads of flies attracted to human waste, aid the dissemination of the disease. In some ways trachoma is more of a social than a medical problem; if living standards can be improved, overcrowding reduced, flies discouraged, and adequate water supplies ensured, the incidence of trachoma decreases rapidly.

The early symptoms of infection are pain, watering of the eye, and sensitivity to light. At this stage the conjunctival lining of the lids is red and velvety in appearance, and the cornea shows gray areas. Later, the conjunctiva appears to have grains of sand embedded in its tissue. Blood vessels grow into the cornea, which becomes thickened and hazy. Secondary bacterial infections are common, but the real dangers of trachoma lie in the scarring and contracture of tissue that occur when healing takes place. These changes affect the upper lid in particular, causing it to buckle inward so that the lashes rub across the already diseased cornea, and it is the corneal scarring thus produced that can cause blindness.

Degenerative conditions. Exposure to wind and dust frequently causes degenerative changes in the exposed part of the conjunctiva, particularly in older people. A yellow nodule forms, first on the nasal side of the cornea and later on the other side. It is without blood vessels and is frequently unnoticed until an incidental conjunctivitis causes it to stand out clearly against the red background of dilated conjunctival vessels. It causes no symptoms and requires no treatment.

A more serious degeneration is that known as a pterygium, found particularly in people who live in hot, dusty climates. It appears as a fleshy growth at the edge of the cornea, with a tendency to progress across its front surface, where it may interfere with vision. Treatment consists of surgical removal, but recurrences are common.

The cornea and sclera. The cornea is the clear window of the eye and its most important refractive surface. Any surface irregularity, any scar in the substance of the cornea, is likely to have a profound effect on vision. Almost the whole nerve supply of the cornea consists of nerve fibres sensitive to pain, so that corneal diseases are always painful and elicit a flow of tears by a reflex action that is part of the protective system of the eye.

Infection of cornea with herpes virus; bacteria; fungi

Inflammation of the cornea. As with inflammations of the conjunctiva, bacterial infection of the cornea has become much less common and viral infections are increasingly important. Of these, the herpes viruses, which cause the common "cold sore" of the lips and skin and the venereal form of herpes, are the most frequent cause of corneal ulceration. Infection is spread by personal contact. The herpes virus causes a typical ulcer of the cornea called, from the pattern of the lesion, dendritic ("branching"). The disease starts with an acutely painful eye, with tearing, and sensitivity to light. The ulcer may heal spontaneously or after medical treatment, but the virus often lies dormant in the tissues; recurrences are common, and with each recurrence there is more danger that the virus will extend deeper into the cornea and cause an intractable inflammation.

Application of the drug deoxyuridine (5-iodo-2-deoxyuridine) to the cornea causes the ulcer to heal more rapidly and reduces the recurrence rate. The action of the drug depends upon its limiting the multiplication of the virus by interfering with the formation of virus deoxyribonucleic acid (DNA) in the host cell.

Bacterial infections of the cornea still occur, usually after injury to the corneal surface, as few bacteria have the power to penetrate the intact surface layers of the cornea. Such ulcers may be extremely severe, and there is always a danger of perforation of the eye, particularly in debilitated patients.

The spores of fungi are commonly present in the atmosphere. The normal cornea is resistant to infection by these organisms, but a fungal infection of the cornea can develop after a corneal injury or other lesion, particularly if corticosteroid drugs have been used in treatment. Intensive treatment with antifungal drugs is usually effective in killing the organisms, but a dense scar is usually left.

A corneal inflammation may start in the deeper layers, usually by spread of infection from the bloodstream. It is seen most commonly in adolescents who have congenital syphilis. Both eyes are usually attacked, although there may be an interval before the second eye is affected. The cornea rapidly becomes hazy, and blood vessels grow in from the surrounding tissues to form a red patch. With the decline in congenital syphilis in developed countries, the condition is now becoming a rarity.

Inflammation of the sclera. The sclera is the fibrous covering of the eye that shows up as a dense white layer beneath the transparent conjunctiva. A relatively mild nodular inflammation sometimes occurs in the superficial layers of the sclera; it is thought to be allergic in nature and usually responds well to anti-inflammatory treatment. Inflammation of the deeper sclera is more severe and often is painful. It occurs more frequently in older people and may be associated with tuberculosis or rheumatism; however, the cause of the condition is often not discovered.

Degenerative conditions. Keratoconus is the name of a curious condition in which the central part of the cornea, normally spherical in shape, begins to bulge and protrude forward as a cone. The only symptom is deterioration of vision due to the irregular astigmatism caused by the changing corneal curvature. Ordinary spectacles cannot correct the irregular refraction, but contact lenses are often of great value, and in more advanced cases corneal grafting is required.

There are numerous other rare types of corneal degeneration, some of which are familial; all produce a deterioration in vision that cannot be corrected with spectacles. Many of these conditions respond well to corneal grafting.

THE INNER EYE

The uveal tract. The uveal tract is a vascular layer of tissue—that is, a layer rich in blood vessels—lying next to the inner surface of the sclera. It is divided into three structures: the choroid, a highly vascular layer that supplies blood to the outer layers of the retina; the ciliary body, largely muscle tissue, which by its contraction and relaxation alters the focussing of the lens; and the iris, the coloured part of the eye, which forms the variable aperture of the eye, the pupil. The ciliary body, which lies at the base of the iris, also functions by forming the aqueous humour, the production and drainage of which regulate intraocular pressure; the aqueous humour also is the source of nutriment to the lens and cornea, which are avascular (without blood vessels).

Inflammation. Inflammations of the uveal tract are always potentially serious because of the secondary effects they may have on the retina and the lens. In most cases the disease affects either the anterior part of the uvea—that is, the iris and ciliary body—or the posterior part, the choroid. An attack of acute anterior uveitis starts with pain, redness, and mistiness of vision. The eye is sensitive to light, and, although there is no discharge as in conjunctivitis, the eye may water. The pupil tends to contract and the normally clear iris markings become less distinct. In chronic anterior uveitis the main symptom is blurring of vision. Acute choroiditis starts with sudden onset of blurring of vision with many black spots floating about in front of the sight.

Except for cases in which the uveitis follows a perforating injury or a corneal ulcer, it is believed that the inflammation is caused by an infective process within the body or by some other mechanism associated with systemic disease. Many infective conditions and parasitic diseases are known to cause uveitis. In a large proportion of cases, however, particularly when the inflammation is confined to the anterior segment, it proves impossible to be sure of the cause in any individual instance, and the investigation of a case of uveitis often poses one of the biggest problems in ophthalmology today. In men, a proportion of cases of anterior uveitis are associated with ankylosing spondylitis, a chronic disease of the joints of the spine, the cause of which is still obscure. Another association, again in men, is with Reiter's disease, a condition that starts as an infection of the urogenital tract with the later development of joint changes, particularly in the sacroiliac joints of the lower back, and recurrent attacks of anterior uveitis. The organism that is responsible for this venereally contracted infection is still unknown, but may be a virus. Infections in the teeth and tonsils have long been held to be a cause of uveitis, and eradication of dental decay does occasionally have a favourable effect on the course of the disease.

Toxoplasmosis and its effect on fetus

Inflammations of the choroid—the posterior portion of the uveal tract—and the retina are more likely to be infective in origin. One of the organisms most commonly involved is *Toxoplasma gondii,* a protozoon of worldwide distribution among domestic animals, small mammals, and man. Although antibodies to the organism can be found in a high proportion of most populations, showing that infection is widespread, overt signs of disease are rarely seen; most people can acquire the infection without being aware of any systemic disturbance at all, and only in special circumstances does the organism cause disease. One of these special circumstances is pregnancy. If a woman becomes infected during pregnancy there is a short period in which invasion of the tissues takes place before circulating antibodies are formed by the mother. During this period it is possible for the organism to pass through the placenta and infect the unborn child. Fetuses appear to be particularly susceptible to the organism, nearly half of those exposed showing some evidence of infection with toxoplasmosis. In severe cases the child may be stillborn or may be born with congenital toxoplasmosis, a serious disease affecting many organs of the body and particularly the brain and eye. In less serious cases small foci of infection are left in the nervous system and the retina of the eye; these may not be apparent at birth and may remain quiescent, only to become active 15 or 20 years later in the form of an inflammation of the choroid and the retina. Children of subsequent pregnancies are unaffected.

The treatment of uveitis has been transformed by the advent of corticosteroid drugs. Even when a specific infection cannot be discovered and treated with the appropriate specific drug, therapy with corticosteroids is usually successful in controlling the worst ravages of the inflammation.

Tumours of the uveal tract. Pigmented tumours are the commonest tumours involving the uveal tract. They may be benign (the nevus or mole) or malignant. The choroid is the commonest site of these lesions, which push the retina forward and cause a retinal detachment. Disturbances of vision are the commonest symptom, but the tumour if neglected may enlarge and cause inflammation and raised pressure within the eye. Small portions of the tumor often enter the bloodstream and settle in distant organs, particularly the liver. The growth of these secondary deposits is often slow; they may not be apparent until many years after the diagnosis of the tumour in the eye.

The lens. The lens is a transparent, avascular organ surrounded by an elastic capsule. It lies behind the pupil and is suspended from the ciliary body by a series of fine ligaments. Its transparency is the result of the regular arrangement of the lens fibres; since these are being formed continuously, the lens continues to grow throughout life. Interference with this growth pattern will result in the formation of abnormal lens fibres that cannot transmit light as well as the normal lens fibres. A small opacity is thus seen in the lens. Minor irregularities are common in otherwise perfectly normal eyes. If the opacity is severe enough to affect vision it is called a cataract.

Congenital lens opacities of many varieties have been recognized and described since the early days of ophthalmology, but they remained curiosities until the work of an Australian ophthalmologist, Norman M. Gregg, threw new light on their cause, and, indeed, on that of many other congenital defects. In 1941 Gregg noticed that after an epidemic of German measles (rubella) many of the children whose mothers had contracted the disease in the first two months of pregnancy were born with cataract, sometimes associated with deafness and congenital heart disease. It is now known that the virus can be recovered from the lens for several months after birth.

Cataract in newborn child; in adult

Cataract in the adult may be the result of injury to the lens by a perforating wound, by exposure to radiation such as X-rays, or as the result of the ingestion of toxic substances or even of some drugs. The lens relies for its nutrition on the aqueous humour secreted by the ciliary body and, if the latter is severely damaged as the result of long-continued uveitis or a tumour, the metabolism of the lens suffers and a cataract develops. The commonest form of cataract is senile cataract, so called because it becomes progressively more common with advancing age. In spite of a large amount of work on the biological and biochemical changes that take place in the lens, the underlying cause of senile cataract is still unknown. Whatever the underlying biochemical changes may be, they result in an increasing clouding of the lens until the whole lens loses its normal transparency and becomes white and opaque. The only symptom is progressive diminution of vision. In the early stages of the condition some visual improvement can usually be obtained with spectacles, but, as the cataract progresses, the visual deterioration becomes sufficiently severe to warrant surgical treatment.

With modern techniques cataract extraction can be done as soon as the visual deterioration interferes with normal activities, and it is no longer necessary for patients to wait for many years in semiblindness to allow the cataract to become mature. Cataract extraction is one of the most successful and satisfying operations in ophthalmic surgery; if the eye is otherwise normal the visual results are excellent, although the refractive power of the lens has to be replaced by a rather thick spectacle lens or special contact lens.

The retina. Developmentally, the retina is part of the

brain and as such has only a limited capacity for repair of its damaged tissue. In particular, the highly specialized rods and cones (the photoreceptors), which are the structures sensitive to light, and the nerve cells of the retina, like those of the brain, cannot be replaced if they are damaged. Death of these cells inevitably has a permanent effect on vision.

The retina is a thin transparent membrane that lines the inner eye. Its outermost layer, the pigment epithelium, is formed of pigmented cells that are closely adherent to the underlying blood vessels of the choroid. The layer of rods and cones is more loosely attached to the pigment epithelium and has complicated nervous networks that culminate in the innermost layer of nerve fibres. These fibres run back through the optic nerve to the brain. The inner two-thirds of the retina derives its blood supply from a special complex of vessels that enters the eye through the optic nerve.

Detachment of retina

Retinal detachment. A retinal detachment is a condition in which the main part of the retina becomes separated from the pigment epithelium. This may follow an injury to the eye or a tumour; or inflammation of the underlying choroid. The commonest type of detachment, however, has no such predisposing factors: the distinctive feature is the formation of a small hole or tear in the retina, usually at the periphery where the retina is thinner. In most cases the hole is caused by an adhesion forming between the retina and the jelly-like substance called the vitreous humour that fills the interior of the eye. Sudden movement of the eye, or an injury, causes the vitreous to pull on the retina, thus creating a tear or hole. When this has happened, fluid can pass through the hole and strip the retina off the pigment epithelium. Myopic (near-sighted) eyes are particularly prone to retinal detachment because they are larger than normal, and the coats of the eye are thinned and stretched. The periphery of the retina, in particular, often shows weak areas, and the vitreous is usually unduly thin and fluid.

The history is often quite typical. The pull of the vitreous on part of the retina creates a sensation of light noticed by the person affected as flashes that occur on movement of the eye. When an actual tear has developed, the retina starts to become detached and the person has the sensation of a shadow coming down over the vision.

The essential factor in treatment is to seal off the hole in the retina. The part of the retina containing the hole must be brought into close contact with the choroid and then by means of a gentle inflammatory reaction caused by using heat, cold, or intense light, the retina is made to stick to the underlying choroid and seal off the leak. The remaining fluid can then be drained away, allowing the retina to fall back into place. Provided that the detachment has not been of long standing, the retinal function recovers quite well once the retina is reattached. The small central area of retina, however, that subserves the most acute vision has only one source of blood supply, the underlying choroid; once it is separated from this some permanent damage ensues, even if the retina is subsequently replaced in its correct position. Thus, it is most important therefore that retinal detachments be treated early, before the central area of the retina becomes detached.

Retinal degeneration. Cases of retinal degeneration can be grouped in two broad classes: hereditary and genetic, and senile. A large number of genetically determined degenerations of the retina have been described. Although they are quite rare, the bizarre appearances of the retina and the inexorable advance of the disease have excited considerable interest among ophthalmologists. These conditions are typified by the disease known as retinitis pigmentosa, a hereditary condition. The earliest symptom is night blindness, which may first be noticed in childhood and is due to alteration in the function of the rods, which are the visual receptors used in dim light. The more peripheral parts of the retina are affected first, and while central vision may be good the field of vision progressively decreases until only a small "tubular field" remains. Cause of the disease is unknown. It is easily recognizable by the narrowing of retinal vessels and the scattering of clumps of pigment throughout the retina.

In senile degeneration, unlike the hereditary type, it is the central part of the vision that is first affected. The central part of the retina, known as the macula, derives its blood supply only from the choroid, and it is probably for this reason that it is likely to suffer first from the slowing of the metabolic changes and from the deficiency of circulation that occur in old age. While degeneration of the macula does not cause blindness, in the sense that the person affected is unable to see anything, it is extremely disturbing because it affects central visual acuity and makes reading or fine work difficult or impossible. There is as yet no satisfactory medical or surgical treatment, but considerable improvement can be obtained by the use of special magnifying spectacles.

The retinal changes that may occur in diabetes, arteriosclerosis, and vascular hypertension are described in a later section.

The optic nerve. The optic nerve, which carries about one million nerve fibres, leaves the globe from the back of the eye and passes through the apex of the orbit into the cranial cavity. It is surrounded by an extension of the membranes that surround the brain and this connection with the intracranial cavity is of some importance, because in some intracranial diseases the pressure within the skull rises and is transmitted along the sheaths of the optic nerve to cause swelling of the optic nerve head, which is visible inside the eye. This swelling of the nerve head, or papilledema, is one of the most important signs of increased intracranial pressure. If the swelling persists, damage to the fibres of the optic nerve takes place, with subsequent loss of vision.

Causes and effects of swelling of optic nerve head

Swelling of the optic nerve may also be caused by inflammatory changes in the nerve itself or in the surrounding sheath; this condition is known as optic neuritis. The symptoms are loss of vision in the central part of the visual field and pain on moving the eye. The condition is most common in young adults and may be due to the spread of infection from the adjacent nasal sinuses. The majority of cases, however, are manifestations of multiple sclerosis, a condition in which the sheath of the nerves becomes altered and interferes with the transference of nervous impulses. This may occur in any part of the nervous system, but the optic nerve is a common site, and the lesion is often the first to be noticed by the patient because of the visual symptoms that result from it. The disease is characterized by long periods of remission from symptoms, and after optic neuritis it may be 10 years or more before other signs are apparent. Usually the function of the optic nerve recovers after an attack of optic neuritis, leaving little, if any, visual disturbance, but there is some atrophy of the fibres.

Optic atrophy may follow any serious disease of the retina involving a large amount of destruction of neural tissue. It may also follow damage to the optic nerve within the skull, or the optic chiasma—that is, the place where the optic nerves crisscross, close to the pituitary gland. Tumours of the pituitary gland nearly always compress the optic nerve fibres and cause some degree of atrophy with loss of vision in that part of the visual field subserved by the fibres concerned. Usually it is the fibres on the inner side of the optic nerve and those that cross at the chiasma that are most involved: these fibres supply the retina on the nasal half. This part of the retina receives visual images from the outer part of the visual field, and in pituitary lesions it is common to find that the outer parts of both visual fields are affected.

Certain chemicals and some drugs can also cause optic atrophy: among them are quinine and methyl alcohol. Optic atrophy is most unlikely to follow normal medical doses of quinine, and when it occurs it is usually from the large doses taken to cause abortion. Methyl alcohol (wood spirit or methylated spirits) is broken down in the body to acetyl aldehyde, which is toxic to neural tissue, and the risks of blindness from drinking methylated spirits are high.

Glaucoma. The thin coats of the eye are not sufficiently rigid in themselves to withstand distortion following the pull of the extraocular muscles when the eye is moved. The eyeball is kept rigid by the action of the ciliary body,

which secretes sufficient amounts of the fluid called the aqueous humour to pump up the pressure of the eye to a level above the atmospheric pressure. This fluid is constantly being formed and drains away at the base of the iris through specialized drainage channels. Should these channels become blocked the pressure within the eye rises to abnormally high levels and impedes the entry of blood into the eye. The fibres of the optic nerve where it enters the eye are particularly susceptible to a reduction in blood supply, and if the intraocular pressure remains raised for long some of these nerve fibres will atrophy, causing loss of function of the retina from which they are derived. Glaucoma is the name given to a condition in which the intraocular pressure is raised to abnormal levels. In some persons this is due to other disease within the eye—such as inflammation or a tumour—but most have one of two distinct diseases, chronic simple glaucoma or closed-angle glaucoma.

Chronic simple glaucoma

Chronic simple glaucoma is a common disease that may affect one percent of people in the older age groups. Although the actual cause is not known it is almost certainly due to degenerative changes in the outflow channels for aqueous fluid. It is rare below the age of 40 but after this its incidence increases; in one recent survey it was found to affect 10 percent of those examined over the age of 80. Genetic influences are important and relatives of patients with glaucoma are five times more likely than others to develop the disease.

The symptoms are slight or absent in the early stages. The slow rise in pressure does not cause pain, and the early visual loss is in the peripheral parts of the visual field, affecting central vision only late in the disease. Both eyes are usually involved, although one may be more severely affected than the other. Since vision lost from glaucoma cannot be restored, successful treatment can only prevent further loss of vision. It is of great importance, therefore, that the disease be diagnosed as early as possible. Measurement of the intraocular pressure is of great value in the diagnosis of glaucoma: this is a simple test that can be applied as a screening method for surveys of the normal population.

From H.G. Scheie, *Medical Ophthalmology. Ophthalmologic Manifestations of Systemic Diseases*

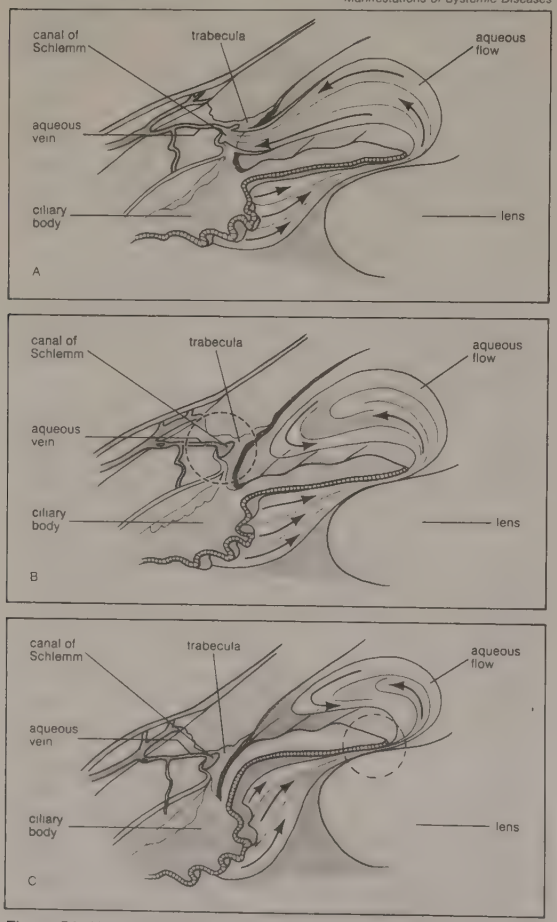


Figure 51: *Normal flow of aqueous humour contrasted with two types of obstruction.*
(A) Normal flow of aqueous humour. (B) Obstruction to flow of aqueous humour in chronic glaucoma. (C) Obstruction to flow of aqueous humour in closed-angle, or acute, glaucoma. Dotted circles indicate site of obstruction.

The medical treatment of chronic simple glaucoma consists of the use of drops that lower the intraocular pressure. Inhibitors of the enzyme carbonic anhydrase, when taken by mouth, reduce the formation of aqueous humour and are used as an additional measure when necessary. If the pressure remains raised in spite of all medical treatment, then surgical methods must be used to increase the drainage of fluid from the eye.

The other common type of glaucoma is called closed-angle glaucoma. This again has a familial incidence and occurs in people who have a rather small, long-sighted eye. Continued growth of the lens in these patients pushes the iris forward and narrows the gap at the root (the outer edge) of the iris where aqueous humour flows out of the eye. This fluid is formed in the ciliary body behind the iris and flows forward through the pupil to the angle of the anterior chamber. The lens pushing against the iris acts as a valve and impedes the flow of aqueous humour through the pupil. The root of the iris, which is rather thin, is then pushed forward and may eventually completely close the exit for aqueous humour, so that the intraocular pressure rises rapidly. The eye becomes painful and the vision is lost; the pain may be so severe as to cause vomiting and prostration. The eye becomes red and stony hard to the touch. Urgent treatment is required to lower the pressure and prevent strangulation of the blood vessels entering the eye.

In some cases an acute attack such as this heralds the onset of the disease, but more frequently minor, subacute, attacks, which are relieved by rest and sleep, occur for months or years. Modern methods of medical treatment are usually effective in lowering the pressure in the acute attack; an operation is usually necessary to prevent further recurrences.

OCULAR INJURIES

The bony orbit provides excellent protection for the eye from blunt injuries. A blow from in front with a rounded instrument such as a fist or tennis ball, however, can cause a shock wave to travel through the eye and damage the retina at the back of the eye. Central vision may be reduced after such injuries without any very obvious changes in the appearance of the eye. In severe cases the bones of the orbit may be fractured. Perforating wounds from glass, sharp metal fragments, and so on, are always serious. Injuries to the lens will result in the formation of a cataract, and often after penetrating injuries the eye remains inflamed for a considerable time.

Sympathetic ophthalmitis

One type of inflammation following injury, sympathetic ophthalmitis, is of particular importance; fortunately it is now rarely seen. The sequence of events is that an injured eye remains irritable and after some weeks, months, or even years, the fellow—previously normal—eye may take part in the inflammation. This is the "sympathizing eye." The cause of sympathetic ophthalmitis is not known, but it is known that if an injured eye is removed within 10 days sympathetic ophthalmitis never occurs in the other eye. In the past there was little effective treatment for the condition, but therapy with corticosteroid hormones has proved effective in controlling the inflammation in most cases, so that even if the disease becomes established the consequences are not as serious as they were previously.

Foreign bodies. Most foreign bodies that enter the eye remain near the surface. When they touch the cornea they cause intense pain and a flow of tears. The tears may be sufficient to wash the foreign body out of the eye, but if it becomes embedded in the cornea it may have to be removed surgically. Many small foreign bodies lodge in the under surface of the upper lid so that every time the eye blinks the foreign body rubs on the cornea, causing pain and irritation.

Small foreign bodies travelling at high speeds may penetrate the interior of the eye with remarkably few symptoms, and their presence may not be recognized until weeks or months later when inflammatory changes occur. The commonest foreign bodies to enter the eye in this way are fragments of metal from hammer-and-chisel accidents, or from moving parts of machinery. Whenever such injuries are suspected, it is important to locate the position of the fragment as carefully as possible, and to remove it by surgery. If the foreign body is magnetic, a large electromagnet is invaluable in attracting the foreign body to the site of the incision in the eye, making extraction comparatively simple.

Chemical and radiation injuries. Strong acids and alkalis always cause severe injury if they enter the eye. Speed is the vital factor in first-aid treatment, and copious irrigation with water the first essential. Delay of first-aid treatment in the hope of finding a neutralizing substance is a serious error, as strong acids and alkalis quickly become bound to the ocular tissues and cause severe necrosis (death of tissue).

Effects of ultraviolet and infrared rays

Except for extremely intense light such as that from a laser, the visible wavelengths of the electromagnetic spectrum—*i.e.*, visible light rays—rarely cause ocular injury. Ultraviolet light, however, is strongly absorbed by the cornea and is the cause of the not uncommon condition known as snow blindness. Symptoms, consisting of intense pain and copious flow of tears, may not occur until some time after exposure. Exposure to light is painful. The treatment consists of cold compresses to the eye and soothing lotions; usually the eyes recover without any permanent damage.

Long-continued exposure to infrared radiation, without adequate protection for the eyes, can cause cataract formation. The lens is also susceptible to X-rays, and the eyes must be shielded when therapeutic irradiation is used for growths around or near the eye. High-intensity microwaves such as those used in some military applications can also cause ocular damage. The widespread use of lasers in research departments and in industry has created a new ocular hazard, and a few cases of accidental exposure have already been reported.

MANIFESTATIONS OF SYSTEMIC DISEASE

The central nervous system. Since the optic nerve and retina are, embryonically, an extension of the brain, it is not surprising that central nervous system diseases frequently affect the eye, and visual defects may be the earliest evidence of the general disease. The nerve supply to the ocular muscles, particularly the extraocular muscles, may also be involved early in some diseases of the central nervous system: this will result in defective movement of the eyes, causing lack of coordination between the two eyes and diplopia, or double vision.

The nerve fibres that connect the retina with the site of visual sensation in the occipital cortex—*i.e.*, the outer brain substance at the back of the head—are arranged throughout the brain in a regular pattern, and many lesions of the brain, such as tumours, impinge on part of this pathway. From a detailed examination of the sensitivity of the different parts of the retina, using tests of the visual field, it is often possible to localize the exact site of an intracranial lesion. An optic neuritis causing sudden onset of loss of central vision in one eye is a frequent first symptom of multiple sclerosis. Detailed ophthalmic examination is therefore essential in the case of any patient suspected of having disease of the central nervous system.

Arteriosclerosis and vascular hypertension. The eye is the one structure in the body in which the blood vessels are easily visible to the examiner, and the changes that can be observed in the retinal vessels mirror those that are taking place in other parts of the body, particularly those in the brain. In arteriosclerosis degenerative changes occur in the walls of the arteries: this leads to thickening of the walls and narrowing of the bloodstream and may give rise to complete occlusion, or blockage of the vessel. If the central retinal artery is affected, loss of vision is complete and sudden and, unless the obstruction can be relieved within an hour or so, permanent. Occlusion of the retinal veins is more common than arterial occlusion and also has dramatic effects: the damming up of blood in the eye results in the bursting of small vessels, and multiple hemorrhages are scattered all over the fundus (that part of the inner eye which can be inspected through the pupil). Some degree of recovery of vision is usual but depends on whether a branch of the central vein or the vein itself is occluded.

Vascular hypertension, or raised blood pressure, usually occurs in association with arteriosclerosis. Typical changes can be recognized in the small vessels of the fundus, and in severe cases multiple hemorrhages and exudates, with swelling of the optic disk (the head of the optic nerve), may be present.

Diabetes. The satisfactory control of diabetes with insulin has increased the incidence of eye complications, for it has become apparent that it is the duration of the disease rather than its severity that determines the onset of ocular changes. A special type of cataract may occur in young diabetics with severe untreated disease, but the most serious complication involves the blood vessels of the retina. The actual cause of the changes in the retinal vessels is still unknown, but the natural history of the disease is well recognized. The retinal capillaries dilate at weak points in the vessel wall—*i.e.*, form small aneurysms; these weak portions of the vessel wall may give way and cause hemorrhages into the retina. In the later stages the hemorrhages become more extensive and spread into the vitreous. New vessels grow into hemorrhagic areas and are followed by fibrous changes that may pull on the retina and cause detachment. Extensive changes of this nature lead invariably to blindness.

Destruction of the pituitary gland, either by direct surgery or by the implantation of a radioactive material, has given some hope of alleviating these severe retinal changes. The procedure is, however, a drastic one. Destruction of affected areas of the retina by the use of an intense beam of light, a process called photocoagulation, promises to be a useful form of treatment in selected cases. Degenerative changes in the retina remain the most serious complication of diabetes.

Thyroid disease. The staring appearance of persons suffering from thyrotoxicosis, also called exophthalmic goitre or Graves' disease, is believed to be due to the stimulation of smooth muscle in the lids and orbit, causing the lid to retract a little from the globe and the globe itself to advance forward slightly. These changes normally regress if the thyrotoxicosis is treated. There is a more serious ocular complication of thyroid disease, which may follow excision of the thyroid for thyrotoxicosis or may, in some cases, arise in persons with normal or subnormal thyroid activity. It is characterized by swelling of the orbital tissues, including the extraocular muscles, so that the eyes cannot be moved properly and project forward between the lids to such an extent that the cornea becomes permanently exposed; the cornea may then ulcerate and even perforate and cause loss of the eye. Sewing the lids together may be sufficient to protect the cornea, but in many cases surgery to relieve the pressure in the orbit is necessary.

Rheumatism. The ocular complications of rheumatoid arthritis mainly involve the sclera, patches of inflammation occurring under the conjunctiva in the scleral and episcleral tissues (the latter are connective tissues between the conjunctiva and the sclera). Although the condition may respond to treatment with corticosteroids, recurrences are common.

VISUAL DISORDERS

"Floaters"; blind areas; flashing lights

Subjective symptoms. One of the commonest visual symptoms is the sensation of small, black objects floating in front of the eye. These move with the eye but lag slightly at the beginning of an eye movement and overshoot when the movement stops. They are due to cells and fragments of debris in the vitreous cavity of the eye. In certain conditions, as when looking at an empty sky, almost everybody can perceive them, and they are normal phenomena. A sudden increase in their number may indicate degenerative changes in the vitreous, which are particularly likely to occur in shortsighted eyes and

in older people. These changes, although annoying, are of no serious import. The appearance of many "floaters," however, may be associated with inflammation or bleeding in the eye.

Blind areas in the field of vision occasionally force people to seek medical advice. Any condition that causes failure of function of part of the retina, the optic nerve, or the optic pathway to the brain, can cause such a blind spot, and the symptom requires careful investigation. There is a naturally occurring "blind spot" in each visual field that corresponds with the lack of retinal elements where the optic nerve enters the eye. The brain is so skillful in filling in the visual pattern that the normal blind spot can be detected only by special methods.

Flashing lights in the field of vision are caused by stimulation of the retina by mechanical means. Most commonly this occurs when the vitreous becomes degenerate and fluid and pulls slightly on its peripheral attachment to the retina. Similar symptoms also arise when the retina becomes detached, causing flashing lights to be seen.

Night blindness and defects of colour perception. Defective vision under reduced illumination may be a rare congenital condition or may be acquired as a result of severe deficiency of vitamin A.

Defective colour vision affects about four percent of men and 0.4 percent of women. Total colour blindness is extremely rare and is nearly always associated with poor vision in ordinary light. The colour-defective person is rarely aware of his disability until special matching tests are used, when it is discovered that he is unable to distinguish between hues in one or another part of the visual spectrum. Other visual functions are perfectly normal, and the only disadvantage is the restriction of certain types of occupation.

Eyestrain. Eyestrain, or asthenopia, is the term used to describe symptoms of fatigue and discomfort following the use of the eyes. Although such symptoms may result from intensive close work, particularly if this is unaccustomed, in people with perfectly normal eyes, they may indicate abnormalities of muscle balance or refractive errors. Eyestrain is more likely to be manifest during periods of fatigue or stress and is common among students working for examinations. Refractive errors require correction and muscle imbalance treatment. Psychological factors are often more important than physical factors.

Presbyopia; short sight; long sight

Refractive errors. In a normal eye rays of light from distant objects come to a focus on the retina. In near vision, the refractive power of the eye is increased by altering the shape of the lens to focus the image on the retina. A twelve-year-old can focus on an object four inches away from the eye but, with age, the ability of the lens to alter its shape decreases so that at the age of 40 the shortest distance at which an object can be kept in focus is about 10 inches. The near point continues to recede with age until fine print, for example, cannot be read at a normal reading distance. This condition is known as presbyopia; it is corrected by the use of convex lenses for reading.

In some eyes rays of light from distant objects are not brought to a focus on the retina but are focused on a plane in front of the retina, as in myopia (short sight), or behind the retina, as in hypermetropia (long sight). In myopia, near objects are brought into focus on the retina but distant objects can only be seen clearly with the aid of concave lenses. In hypermetropia, distant objects can usually be brought into focus by using the accommodative power of the lens, and in young people there is usually sufficient accommodation to enable them to see reasonably near to them. The constant accommodative effort required, however, may produce symptoms, and the necessity for accommodating for distance can be overcome by wearing convex glasses.

Another type of refractive error is astigmatism. In this condition the refractive power of the eye varies in different axes because of variation in curvature so that vision at all distances is distorted and can only be corrected by the use of cylindrical lenses or contact lenses.

Minor degrees of refractive error are extremely common. The refractive state is genetically determined and there are marked racial differences. Myopia, for example, is common in the Far East and rare in the African Negro. Although most refractive errors are easily correctible by spectacles and such errors are rarely accompanied by any serious disease of the eyes, hypermetropia is a factor in the development of some kinds of squint and high degrees of myopia are often associated with serious degenerative changes within the eye.

OPHTHALMOLOGICAL EXAMINATION AND CORRECTIVE DEVICES

Ophthalmological examination. An ophthalmological examination comprises a history of a patient's symptoms and signs, subjective tests to determine the visual function, and physical examination of the eyes by means of special devices. The most important subjective test is for visual acuity, and this is usually performed by presenting to the patient a series of letters of graded sizes at a set distance. He is required to read the lowest line legible to him; visual acuity can then be expressed in terms of the size of the letter and the distance at which it is read.

The visual field is assessed by moving an illuminated target inward from the periphery toward a central point viewed by the eye: the area in which the target is seen can then be drawn as a map of the visual field for that eye.

Other subjective examinations include colour-vision testing and tests of visual perception under reduced illumination. Examination of the external eye and part of the anterior segment is facilitated by the use of a binocular microscope mounted horizontally, to which is attached a slit-lamp, a variable source of light that projects the image of a slit onto the eye. The ophthalmoscope has an illuminating system that lights up the interior of the eye and a viewing system through which the fundus can be observed. Photography of the anterior part of the eye and of the fundus is both possible and widely used.

Other specialized methods of examination include examination of the angle of the anterior chamber by means of a specially designed contact lens with the slit-lamp microscope. The electrical responses of the retina and brain to light entering the eye can also be recorded and are of great value in certain conditions.

Estimation of the intraocular pressure is an important part of an ophthalmological examination and is accomplished by an instrument called a tonometer. This instrument is designed specifically to measure the tension or pressure that exists within the eyeball.

The refractive state of the eye can be measured objectively, or subjectively, or by a combination of both methods. The simplest method is subjective, using lenses of different powers to give a trial-and-error estimate of the best correcting lenses. More accurate results can be obtained by using an instrument known as a retinoscope, which gives an objective assessment of the refraction that can subsequently be modified by subjective methods to suit the individual requirements of the patient.

Types of spectacles

Optical aids. The most widely used optical aids are spectacles, and the technical design of spectacle lenses has advanced considerably in the last 50 years. A simple biconcave or biconvex lens causes considerable distortion of appearances if objects are viewed through the periphery of the lens, but if the back surface of the lens is made concave and the required power attained by altering the curvature of the front surface, improvement in peripheral definition results. All modern spectacle lenses are of this form.

Most older people require an additional lens for reading, and this can be incorporated with the distance correction in the form of a bifocal lens. In some occupations an intermediate distance is also required, and a third segment can be added, forming a trifocal lens. The complete range of correction from distance to near can only be achieved by means of a multifocal lens, and these are now available: the upper segment provides the correction for distance; as the eye moves lower down the lens its power increases, the lowest segment of the lens representing the reading correction. By slightly tilting the head it is possible to find the optimum correction for any intermediate distance.

The distortion of peripheral view when using conventional spectacles occurs because the correcting lens does not move when the eye moves. This problem can be

completely overcome by the use of contact lenses, which fit the anterior surface of the cornea and thus move with the eye. The earliest types were larger than the cornea and were uncomfortable to wear, but the smaller "hard" lens greatly increased the scope and usefulness of contact lenses. Even so, the length of time for which they could be worn was limited. Further advances in design and materials—for example, the flexible "soft" lenses—extended the use of contact lenses even further. Made of water-absorbing plastic gel, soft lenses are usually more comfortable for the wearer because they allow oxygen to reach the surface of the cornea and are less likely to scratch or irritate it. A further advance is the disposable lens, a type of thin soft lens that can be worn for only one day or one week.

For those persons who cannot obtain useful vision with ordinary spectacles or contact lenses, much can still be done by the use of compound lens systems known as low vision aids. These devices provide a magnified image but inevitably reduce the visual field. Their main value is to enable a person to read normal print who would otherwise be unable to read. They can be of use for distance, particularly when viewing conditions are relatively static, as with the cinema, theatre, or television.

Finally, for those who are completely blind from ocular causes, there is new hope in the development of implants into the visual cortex that can be connected to a small television camera in such a way that electrical signals can be applied to the visual cortex, completely bypassing the normal optic pathways. The miniaturization of electrical circuitry resulting from space research has made the design of such devices a practical possibility. Their application to human subjects, however, is still in the experimental stage.

BLINDNESS

It is difficult to obtain reliable statistical information on the incidence of blindness on a worldwide basis. Even in countries in which the registration of blind people is attempted, the definitions of "blindness" vary from one country to another; in large parts of the world there is no registration, and the only estimate that can be made depends on random surveys of small parts of the population. An incidence of about 200 per 100,000 is fairly representative of countries in which the standard of medical care is high; it is probable that the incidence is 10 times higher in countries in which medical care is rudimentary.

Causes of blindness

There is wide variation in the causes of blindness in different parts of the world. This is partly due to geographic and climatic conditions, but, more important, it is also due to differences in standards of hygiene and the availability of medical care. Infections, particularly trachoma, spread most easily in warm countries where the population is often crowded into small villages with lack of adequate hygienic facilities. Cataract is still high on the list of causes of blindness in many countries in the world, even though it is so easily curable by surgical means. As the standards of general medical care increase and the expectation of life increases, so the pattern of blindness changes and degenerative conditions, diabetic disorders of the retina, and genetically determined diseases become predominant. Advances in the prevention and the medical and surgical treatment of blindness can be of benefit only to a population that has access to medical care. Until the nutritional and hygienic standards of a large part of the world population can be improved, preventable blindness will remain at its present high level. (E.S.P.)

HUMAN HEARING AND BALANCE: STRUCTURE AND FUNCTION OF THE EAR

Parts of the ear

The human ear, like that of other mammals, contains sense organs that serve two quite different functions: that of hearing and that of postural equilibrium and coordination of head and eye movements. Anatomically the ear has three distinguishable parts: the outer, middle, and inner ear (Figure 52). The outer ear consists of the visible portion called the auricle, or pinna, which projects from the side of the head, and the short external auditory canal, the inner end of which is closed by the tympanic membrane, commonly called the eardrum. The function of the outer ear is to collect sound waves and guide them to the tympanic membrane. The middle ear is a narrow, air-filled cavity in the temporal bone. It is spanned by a chain of three tiny bones—the malleus (hammer), incus (anvil), and stapes (stirrup), collectively called the auditory ossicles. This ossicular chain conducts sound from the tympanic membrane to the inner ear, which has been known since the time of Galen (2nd century AD) as the labyrinth. It is

Figure 52: Structure of the human ear.

Encyclopædia Britannica, Inc

a complicated system of fluid-filled passages and cavities located deep within the rock-hard petrous portion of the temporal bone. The inner ear consists of two functional units: the vestibular apparatus, consisting of the vestibule and semicircular canals, which contains the sensory organs of postural equilibrium; and the snail-shell-like cochlea, which contains the sensory organ of hearing. These sensory organs are highly specialized endings of the eighth cranial nerve, also called the vestibulocochlear nerve.

Anatomy of the human ear

OUTER EAR

The most striking differences between the human ear and the ears of other mammals are in the structure of the outermost part, the auricle (Figure 52). In humans the auricle is an almost rudimentary, usually immobile shell that lies close to the side of the head. It consists of a thin plate of yellow fibrocartilage covered by closely adherent skin. The cartilage is molded into clearly defined hollows, ridges, and furrows that form an irregular, shallow funnel. The deepest depression, which leads directly to the external auditory canal, or acoustic meatus, is called the concha. It is partly covered by two small projections, the tongue-like tragus in front and the antitragus behind. Above the tragus a prominent ridge, the helix, arises from the floor of the concha and continues as the incurved rim of the upper portion of the auricle. An inner, concentric ridge, the antihelix, surrounds the concha and is separated from the helix by a furrow, the scapha, also called the fossa of the helix. In some ears a little prominence known as Darwin's tubercle is seen along the upper, posterior portion of the helix; it is the vestige of the folded-over point of the ear of a remote human ancestor. The lobule, the fleshy lower part of the auricle, is the only area of the outer ear that contains no cartilage.

External auditory canal

The external auditory canal is a slightly curved tube that extends inward from the floor of the concha and ends blindly at the tympanic membrane (Figure 52). In its outer third the wall of the canal consists of cartilage; in its inner two-thirds, of bone. The entire length of the passage (24 millimetres, or almost 1 inch) is lined with skin, which also covers the outer surface of the tympanic membrane. Fine hairs directed outward and modified sweat glands that produce earwax, or cerumen, line the canal and discourage insects from entering it.

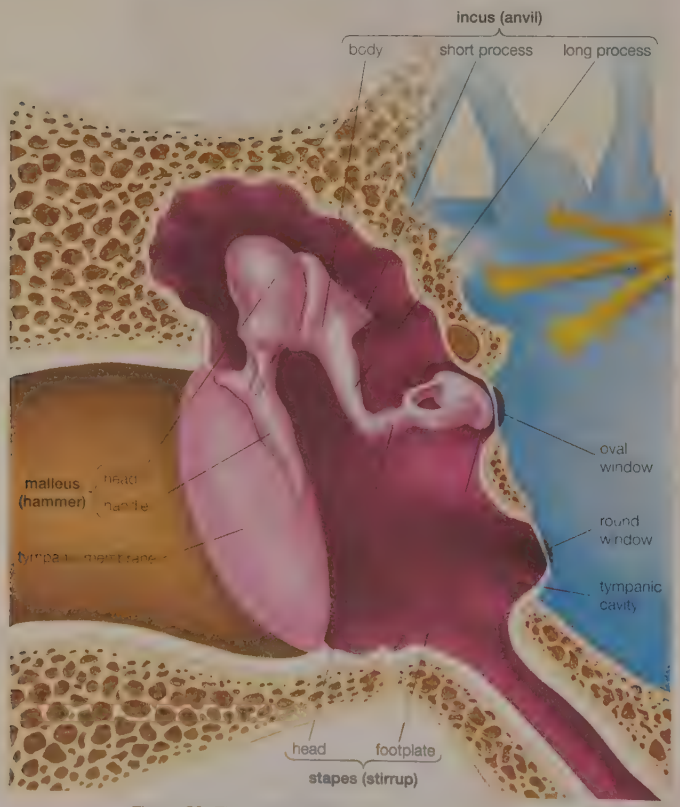


Figure 53: The auditory ossicles of the middle ear and the structures surrounding them.

Encyclopædia Britannica, Inc

TYMPANIC MEMBRANE AND MIDDLE EAR

Tympanic membrane. The thin, semitransparent tympanic membrane, or eardrum, which forms the boundary between the outer and middle ear, is stretched obliquely across the end of the external canal. Its diameter is about 9 millimetres (0.35 inch), its shape that of a flattened cone with its apex directed inward. Thus, its outer surface is slightly concave. The edge of the membrane is thickened and attached to a groove in an incomplete ring of bone, the tympanic annulus, which almost encircles it and holds it in place. The uppermost small area of the membrane where the ring is open is slack and is called the pars flaccida, but the far greater portion is tightly stretched and is called the pars tensa. The appearance and mobility of the tympanic membrane are important for the diagnosis of middle-ear disease, which is especially common in young children. When viewed with the otoscope, the healthy membrane is translucent and pearl-gray in colour, sometimes with a pinkish or yellowish tinge.

Middle-ear cavity. The cavity of the middle ear is a narrow, air-filled space. A slight constriction divides it into an upper and a lower chamber, the tympanum (tympanic cavity) proper below and the epitympanum above. These chambers also are referred to as the atrium and attic, respectively. The middle-ear space roughly resembles a rectangular room with four walls, a floor, and a ceiling. The outer (lateral) wall of the middle-ear space is formed by the tympanic membrane. Its ceiling (superior wall) is a thin plate of bone that separates it from the cranial cavity and brain above. The floor (inferior wall) is also a thin bony plate separating the cavity from the jugular vein and carotid artery below. The back (posterior) wall partly separates it from another cavity, the mastoid antrum, but an opening in this wall leads to the antrum and to the small air cells of the mastoid process, which is the roughened, slightly bulging portion of the temporal bone just behind the external auditory canal and the auricle. In the front (anterior) wall is the opening of the eustachian, or auditory, tube, which connects the middle ear with the nasopharynx (see below *Eustachian tube*). The inner (medial) wall, which separates the middle ear from the inner ear, or labyrinth, is a part of the bony otic capsule of the inner ear. It has two small openings, or fenestrae, one above the other. The upper one is the oval window, which is closed by the footplate of the stapes. The lower one is the round window, which is covered by a thin membrane.

Tympanum and epitympanum

Auditory ossicles. Crossing the middle-ear cavity is the short ossicular chain formed by three tiny bones that link the tympanic membrane with the oval window and inner ear (Figure 53). From the outside inward they are the malleus (hammer), the incus (anvil), and the stapes (stirrup). These bones are suspended by ligaments, which leave the chain free to vibrate in transmitting sound from the tympanic membrane to the inner ear.

The malleus consists of a handle and a head. The handle is firmly attached to the tympanic membrane from the centre (umbo) to the upper margin (Figure 53). The head of the malleus and the body of the incus are joined tightly and are suspended in the epitympanum just above the upper rim of the tympanic annulus, where three small ligaments anchor the head of the malleus to the walls and roof of the epitympanum. Another minute ligament fixes the short process (crus) of the incus in a shallow depression, called the fossa incudis, in the rear wall of the cavity. The long process of the incus is bent near its end and bears a small bony knob that forms a loose, ligament-enclosed joint with the head of the stapes. The stapes is the smallest bone in the body. It is about 3 millimetres (0.1 inch) long and weighs scarcely 3 milligrams (0.0001 ounce). It lies almost horizontally, at right angles to the process of the incus. Its base, or footplate, fits nicely in the oval window and is surrounded by the elastic annular ligament, although it remains free to vibrate in transmitting sound to the labyrinth.

Muscles. Two minuscule muscles are located in the middle ear. The longer muscle, called the tensor tympani, emerges from a bony canal just above the opening of the eustachian tube and runs backward then outward as it changes direction in passing over a pulleylike projection of bone. The tendon of this muscle is attached to the upper part of the handle of the malleus. When contracted, the tensor tympani tends to pull the malleus inward and thus maintains or increases the tension of the tympanic membrane. The shorter, stouter muscle, called the stapedius, arises from the back wall of the middle-ear cavity and extends forward and attaches to the neck of the head of the stapes. Its reflex contractions tend to tip the stapes backward, as if to pull it out of the oval window. Thus it selectively reduces the intensity of sounds entering the inner ear, especially those of lower frequency.

Eustachian tube. The eustachian tube, about 45 millimetres (1.75 inches) long, leads downward and inward from the tympanum to the nasopharynx, the space that is behind and continuous with the nasal passages and is above the soft palate. At its upper end the tube is narrow and surrounded by bone. Nearer the pharynx it widens and becomes cartilaginous. Its mucous lining, which is continuous with that of the middle ear, is covered with cilia, small hairlike projections whose coordinated rhythmical sweeping motions speed the drainage of mucous secretions from the tympanum to the pharynx.

The eustachian tube helps to ventilate the middle ear and to maintain equal air pressure on both sides of the tympanic membrane. The tube is closed at rest and opens during swallowing so that minor pressure differences are adjusted without conscious effort.

INNER EAR

The membranous and bony labyrinths

There are actually two labyrinths of the inner ear, one inside the other—the membranous labyrinth contained within the bony labyrinth (Figure 54). The bony labyrinth consists of a central chamber called the vestibule, the three semicircular canals, and the spirally coiled cochlea. Within each structure, and filling only a fraction of the available space, is a corresponding portion of the membranous labyrinth: the vestibule contains the utricle and saccule, each semicircular canal its semicircular duct, and the cochlea its cochlear duct. Surrounding the membranous labyrinth and filling the remaining space is the watery fluid called perilymph. It is derived from blood plasma and resembles but is not identical with the cerebrospinal fluid of the brain and the aqueous humour of the eye. Like most of the hollow organs, the membranous labyrinth is lined with epithelium (a sheet of specialized cells that covers internal and external body surfaces). It is filled with a fluid called endolymph, which has a markedly different ionic content from perilymph. Because the membranous labyrinth is a closed system, the endolymph and perilymph do not mix.

Vestibular system. The vestibular system is the apparatus of the inner ear involved in balance. It consists of two structures of the bony labyrinth, the vestibule and the semicircular canals, and the structures of the membranous labyrinth contained within them (Figure 54).

Vestibule. The two membranous sacs of the vestibule, the utricle and the saccule, are known as the otolith organs (Figure 55). Because they respond to gravitational forces, they are also called gravity receptors. Each sac has on its inner surface a single patch of sensory cells called a macula, which is about 2 millimetres (0.08 inch) in diameter and which monitors the position of the head relative to the vertical (see below *The physiology of balance: vestibular function: Detection of linear acceleration: static equilibrium*). In the utricle the macula projects from the anterior wall of that tubular sac and lies primarily in the horizontal plane. In the saccule the macula is in the vertical plane and directly overlies the bone of the inner wall of the vestibule. In shape it is elongated and resembles the letter J. Each macula consists of neuroepithelium, a layer that is made up of supporting cells and sensory cells, as well as a basement membrane, nerve fibres and nerve endings, and underlying connective tissue. The sensory cells are called hair cells because of the hairlike cilia—stiff, nonmotile stereocilia and flexible, motile kinocilia—that project from their apical ends. The nerve fibres are from the superior, or vestibular, division of the vestibulocochlear nerve. They pierce the basement membrane and, depending on the type of hair cell, either end on the basal end of the cell or form a calyx, or cuplike structure, that surrounds it.

Each of the hair cells of the vestibular organs is topped by a hair bundle, which consists of about 100 fine, nonmotile stereocilia of graded lengths and a single motile kinocilium. The stereocilia are anchored in a dense cuticular plate at the cell's apex. The single kinocilium, which is larger and longer than the stereocilia, rises from a noncuticular area of the cell membrane at one side of the cuticular plate. The tallest stereocilia are those closest to the kinocilium, and they decrease in length in stepwise fashion away from the kinocilium. Minute filamentous strands link the tips and shafts of neighbouring stereocilia to each other. When the hair bundles are deflected—*e.g.*, because of a tilt of the head—the hair cells are stimulated to alter the rate of the nerve impulses that they are constantly sending via the vestibular nerve fibres to the brain stem. Covering the entire macula is a delicate acellular structure, the otolithic, or statolithic, membrane. This membrane is sometimes described as gelatinous, although it has a fibrillar pattern. The surface of the membrane is covered by a blanket of rhombohedral crystals, referred to as otoconia, or statoconia, and which consist of calcium carbonate in the form of calcite. These crystalline particles, which range in length from 1 to 20 micrometres (there are about 25,000 micrometres in an inch), are much denser than the membrane—their specific gravity is almost three times that of the membrane and the endolymph—and thus add considerable mass to it.

The vestibular hair cells are of two types. Type I cells have a rounded body enclosed by a nerve calyx; type II cells have a cylindrical body with nerve endings at the base. They form a mosaic on the surface of the maculae, with the type I cells dominating in a curvilinear area (the striola) near the centre of the macula and the cylindrical cells around the periphery. The significance of these patterns is poorly understood, but they may increase sensitivity to slight tiltings of the head.

Semicircular canals. The three semicircular canals of the bony labyrinth are designated, according to their position, superior, horizontal, and posterior (Figure 54). The superior and posterior canals are in diagonal vertical

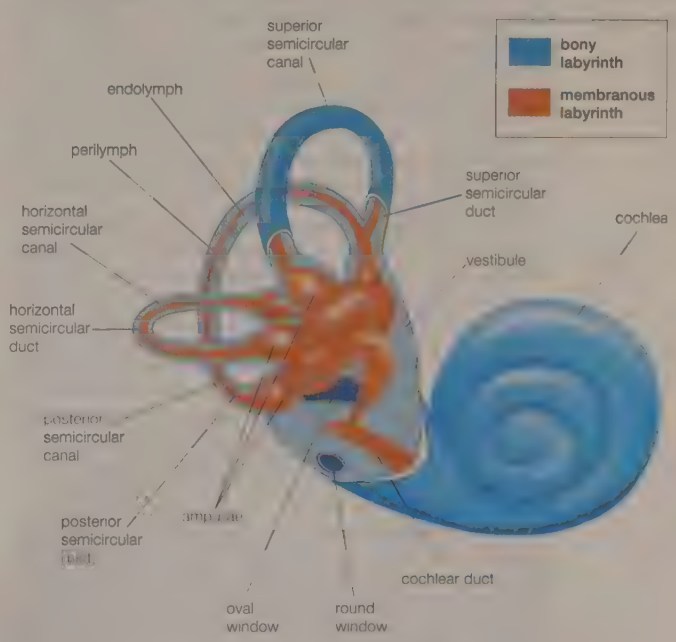


Figure 54: The two labyrinths of the inner ear. The bony labyrinth is partially cut away to show the membranous labyrinth within.

Encyclopædia Britannica, Inc

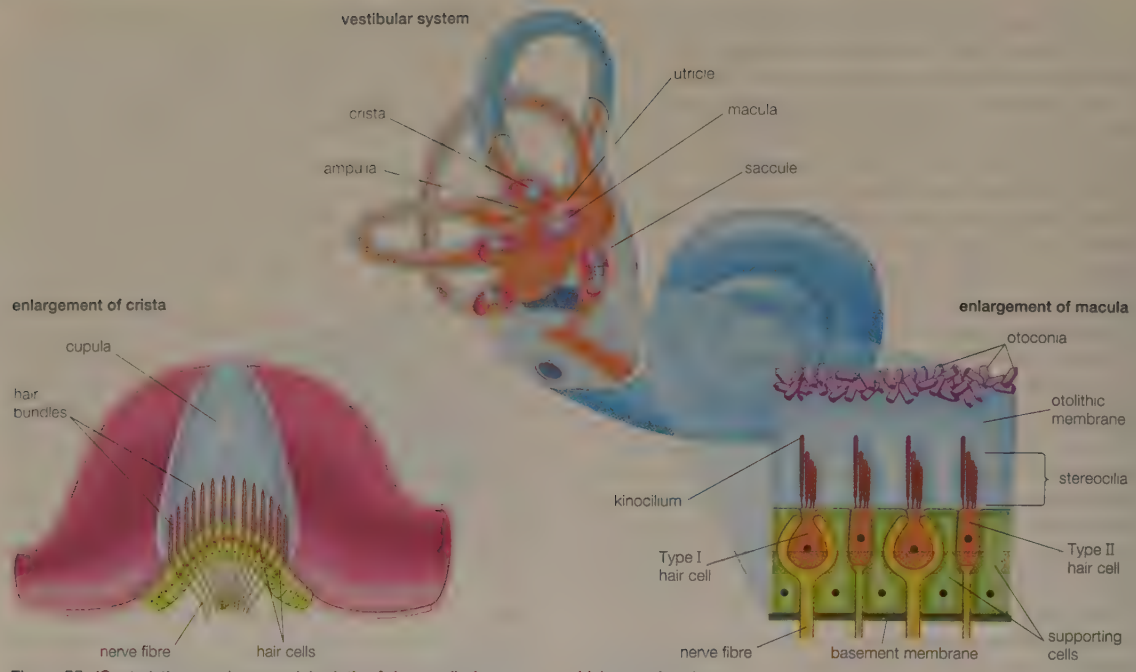


Figure 55: (Centre) the membranous labyrinth of the vestibular system, which contains the organs of balance: (lower left) the cristae of the semicircular ducts and (lower right) the maculae of the utricle and saccule.

Encyclopædia Britannica, Inc.

planes that intersect at right angles. Each canal has an expanded end, the ampulla, which opens into the vestibule. The ampullae of the horizontal and superior canals lie close together, just above the oval window, but the ampulla of the posterior canal opens on the opposite side of the vestibule. The other ends of the superior and posterior canals join to form a common stem, or crus, which also opens into the vestibule. Nearby is the mouth of a canal called the vestibular aqueduct, which opens into the cranial cavity. The other end of the horizontal canal has a separate opening into the vestibule. Thus, the vestibule completes the circle for each of the semicircular canals.

Semicircular ducts

Each of the three bony canals and their ampullae encloses a membranous semicircular duct of much smaller diameter that has its own ampulla. The membranous ducts and ampullae follow the same pattern as the canals and ampullae of the bony labyrinth, with their openings into the utricle and with a common crus for the superior and posterior ducts. Like the other parts of the membranous labyrinth, they are filled with endolymph and surrounded by perilymph. The narrow endolymphatic duct passes from the utricle through the vestibular aqueduct into the cranial cavity, carrying excess endolymph to be absorbed by the endolymphatic sac.

Each membranous ampulla contains a saddle-shaped ridge of tissue called the crista, the sensory end organ that extends across it from side to side. It is covered by neuroepithelium, with hair cells and supporting cells. From this ridge rises a gelatinous structure, the cupula, which extends to the roof of the ampulla immediately above it, dividing the interior of the ampulla into two approximately equal parts. Like the hair cells of the maculae, the hair cells of the cristae have hair bundles projecting from their apices. The kinocilium and the longest stereocilia extend far up into the substance of the cupula, occupying fine parallel channels. Thus, the cupula is attached at its base to the crista but is free to incline toward or away from the utricle in response to the slightest flow of endolymph or a change in pressure. The tufts of cilia move with the cupula and, depending on the direction of their bending, cause an increase or decrease in the rate of nerve impulse discharges carried by the vestibular nerve fibres to the brain stem.

Cochlea. *Structure of the cochlea.* The cochlea contains the sensory organ of hearing. It bears a striking resemblance to the shell of a snail and in fact takes its name from the Greek word for this object. The cochlea is a spiral tube that is coiled two and one-half turns around a hollow central pillar, the modiolus. It forms a cone approximately 9 millimetres (0.35 inch) in diameter at its base and 5 millimetres in height. When stretched out, the tube is approximately 30 millimetres in length; it is widest—2 millimetres—at the point where the basal coil opens into the vestibule and tapers until it ends blindly at the apex. The otherwise hollow centre of the modiolus contains the cochlear artery and vein, as well as the twisted trunk of fibres of the cochlear nerve. This nerve, a division of the very short vestibulocochlear nerve, enters the base of the modiolus from the brain stem through an opening in the petrous portion of the temporal bone called the internal meatus. The spiral ganglion cells of the cochlear nerve are found in a bony spiral canal winding around the central core.

A thin bony shelf, the osseous spiral lamina, winds

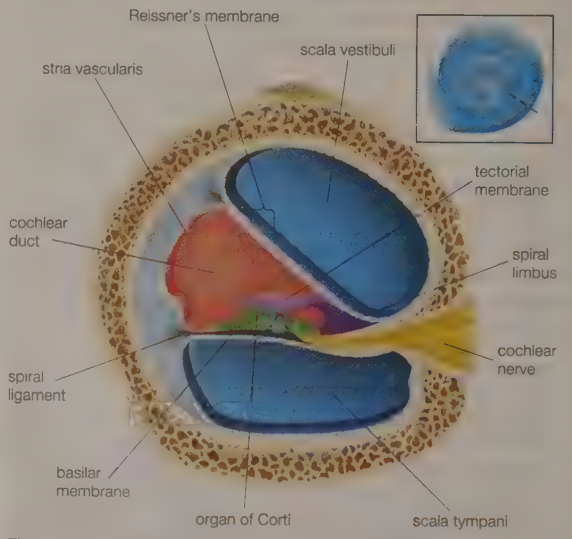


Figure 56: A cross section through one of the turns of the cochlea (inset) showing the scala tympani and scala vestibuli, which contain perilymph, and the cochlear duct, which is filled with endolymph.

Encyclopædia Britannica, Inc.

around the modiolus like the thread of a screw. It projects about halfway across the cochlear canal, partly dividing it into two compartments, an upper chamber called the scala vestibuli (vestibular ramp) and a lower chamber called the scala tympani (tympanic ramp). The scala vestibuli and scala tympani, which are filled with perilymph, communicate with each other through an opening at the apex of the cochlea, called the helicotrema, which can be seen if the cochlea is sliced longitudinally down the middle. At its basal end, near the middle ear, the scala vestibuli opens into the vestibule. The basal end of the scala tympani ends blindly just below the round window. Nearby is the opening of the narrow cochlear aqueduct, through which passes the perilymphatic duct. This duct connects the interior of the cochlea with the subdural space in the posterior cranial fossa (the rear portion of the floor of the cranial cavity).

Cochlear duct

A smaller scala, called the cochlear duct (scala media), lies between the larger vestibular and tympanic scalae; it is the cochlear portion of the membranous labyrinth. Filled with endolymph, the cochlear duct ends blindly at both ends—*i.e.,* below the round window and at the apex. In cross section this duct resembles a right triangle (Figure 56). Its base is formed by the osseous spiral lamina and the basilar membrane, which separate the cochlear duct from the scala tympani. Resting on the basilar membrane is the organ of Corti, which contains the hair cells that give rise to nerve signals in response to sound vibrations. The side of the triangle is formed by two tissues that line the bony wall of the cochlea, the stria vascularis, which lines the outer wall of the cochlear duct, and the fibrous spiral ligament, which lies between the stria and the bony wall of the cochlea. A layer of flat cells bounds the stria and separates it from the spiral ligament. The hypotenuse is formed by the transparent vestibular membrane of Reissner, which consists of only two layers of flattened cells. A low ridge, the spiral limbus, rests on the margin of the osseous spiral lamina. Reissner's membrane stretches from the inner margin of the limbus to the upper border of the stria.

Basilar membrane

In humans the basilar membrane is about 30 to 35 millimetres in length. It widens from less than 0.001 millimetre near its basal end to 0.005 millimetre near the apex. The basilar membrane is spanned by stiff, elastic fibres that are connected at their basal ends in the modiolus. Their distal ends are embedded in the membrane but are not actually attached, which allows them to vibrate. The fibres decrease in calibre and increase in length from the basal end of the cochlea near the middle ear to the apex, so that the basilar membrane as a whole decreases remarkably in stiffness from base to apex. Furthermore, at the basal end the osseous spiral lamina is broader, the stria vascularis wider, and the spiral ligament stouter than at the apex. In contrast, however, the mass of the organ of Corti is least at the base and greatest at the apex. Thus, a certain degree of tuning is provided in the structure of the cochlear duct and its contents. With greater stiffness and less mass, the basal end is more attuned to the sounds of higher frequencies. Decreased stiffness and increased mass render the apical end more responsive to lower frequencies.

Organ of Corti. Arranged on the surface of the basilar membrane are orderly rows of the sensory hair cells, which generate nerve impulses in response to sound vibrations. Together with their supporting cells they form a complex neuroepithelium called the basilar papilla, or organ of Corti. The organ of Corti is named after the Italian anatomist Alfonso Corti, who first described it in 1851. Viewed in cross section (Figure 57) the most striking feature of the organ of Corti is the arch, or tunnel, of Corti, formed by two rows of pillar cells, or rods. The pillar cells furnish the major support of this structure. They separate a single row of larger, pear-shaped, inner hair cells from three or more rows of smaller, cylindrical, outer hair cells. The inner hair cells are supported and enclosed by the inner phalangeal cells, which rest on the thin outer portion, called the tympanic lip, of the spiral limbus. On the inner side of the inner hair cells and the cells that support them is a curved furrow called the inner sulcus. This is lined with more or less undifferentiated cuboidal cells.

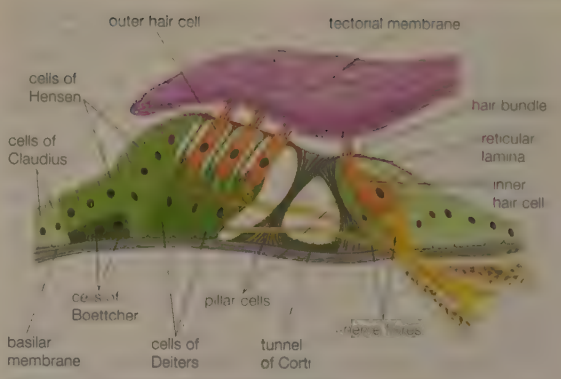


Figure 57: Structure of the organ of Corti.
Encyclopædia Britannica, Inc.

Each outer hair cell is supported by a phalangeal cell of Deiters, or supporting cell, which holds the base of the hair cell in a cup-shaped depression (Figure 57). From each Deiters' cell a projection extends upward to the stiff membrane, the reticular lamina, that covers the organ of Corti. The top of the hair cell is firmly held by the lamina, but the body is suspended in fluid that fills the space of Nuel and the tunnel of Corti.

Each hair cell has a cytoskeleton composed of filaments of the protein actin, which imparts stiffness to structures in which it is found. The hair cell is capped by a dense cuticular plate, composed of actin filaments, which bears a tuft of stiffly erect stereocilia, also containing actin, of graded lengths arranged in a staircase pattern. This so-called hair bundle has rootlets anchored firmly in the cuticular plate. On the top of the inner hair cells 40 to 60 stereocilia are arranged in two or more irregularly parallel rows. On the outer hair cells approximately 100 stereocilia form a W pattern. At the notch of the W the plate is incomplete, with only a thin cell membrane taking its place. Beneath the membrane is the basal body of a kinocilium, although no motile ciliary (hairlike) portion is present as is the case on the hair cells of the vestibular system.

Stereocilia

The stereocilia are about three to five micrometres in length. The longest make contact with but do not penetrate the tectorial membrane (Figure 57). This membrane is an acellular, gelatinous structure that covers the top of the spiral limbus as a thin fibrillar layer, then becomes thicker as it extends outward over the inner sulcus and the reticular lamina. Its fibrils extend radially and somewhat obliquely to end at its lateral border, just above the junction of the reticular lamina and the cells of Hensen. In the upper turns of the cochlea, the margin of the membrane ends in fingerlike projections that make contact with the stereocilia of the outermost hair cells.

The myelin-ensheathed fibres of the vestibulocochlear nerve fan out in spiral fashion from the modiolus to pass into the channel near the root of the osseous spiral lamina, called the canal of Rosenthal. The bipolar cell bodies of these neurons constitute the spiral ganglion. Beyond the ganglion their distal processes extend radially outward in the bony lamina beneath the limbus to pass through an array of small pores directly under the inner hair cells, called the habenula perforata. Here the fibres abruptly lose their multilayered coats of myelin and continue as thin, naked, unmyelinated fibres into the organ of Corti. Some fibres form a longitudinally directed bundle running beneath the inner hair cells and another bundle just inside the tunnel, above the feet of the inner pillar cells. The majority of the fibres (some 95 percent in the human ear) end on the inner hair cells. The remainder cross the tunnel to form longitudinal bundles beneath the rows of the outer hair cells on which they eventually terminate.

The endings of the nerve fibres beneath the hair cells are of two distinct types. The larger and more numerous endings contain many minute vesicles, or liquid-filled sacs, containing neurotransmitters, which mediate impulse transmission at neural junctions. These endings belong to a special bundle of nerve fibres that arise in the brain stem and constitute an efferent system, or feedback loop,

to the cochlea. The smaller and less numerous endings contain few vesicles or other cell structures. They are the terminations of the afferent fibres of the cochlear nerve, which transmit impulses from the hair cells to the brain stem (see below *The physiology of hearing: Cochlear nerve and central auditory pathways*).

The total number of outer hair cells in the cochlea has been estimated at 12,000 and the number of inner hair cells at 3,500. Although there are about 30,000 fibres in the cochlear nerve, there is considerable overlap in the innervation of the outer hair cells. A single fibre may supply endings to many hair cells, which thus share a "party line." Furthermore, a single hair cell may receive nerve endings from many fibres. The actual distribution of nerve fibres in the organ of Corti has not been worked out in detail, but it is known that the inner hair cells receive the lion's share of afferent fibre endings without the overlapping and sharing of fibres that are characteristic of the outer hair cells.

Viewed from above, the organ of Corti with its covering, the reticular lamina, forms a well-defined mosaic pattern. In humans the arrangement of the outer hair cells in the basal turn of the cochlea is quite regular, with three distinct and orderly rows; but in the higher turns of the cochlea the arrangement becomes slightly irregular, as scattered cells form fourth or fifth rows. The spaces between the outer hair cells are filled by oddly shaped extensions (phalangeal plates) of the supporting cells. The double row of head plates of the inner and outer pillar cells cover the tunnel and separate the inner from the outer hair cells. The reticular lamina extends from the inner border cells near the inner sulcus to the Hensen cells but does not include either of these cell groups. When a hair cell degenerates and disappears as a result of aging, disease, or noise-induced injury, its place is quickly covered by the adjacent phalangeal plates, which expand to form an easily recognized "scar."

Endolymph and perilymph. The perilymph, which fills the space within the bony labyrinth surrounding the membranous labyrinth, is similar, but not identical, in composition to other extracellular fluids of the body, such as cerebrospinal fluid. The concentration of sodium ions in the perilymph is high (about 150 milliequivalents per litre), and that of potassium ions is low (about 5 milliequivalents per litre), as is true of other extracellular fluids. Like these fluids, the perilymph is apparently formed locally from the blood plasma by transport mechanisms that selectively allow substances to cross the walls of the capillaries. Although it is anatomically possible for cerebrospinal fluid to enter the cochlea by way of the perilymphatic duct, experimental studies have made it appear unlikely that the cerebrospinal fluid is involved in the normal production of perilymph.

The membranous labyrinth is filled with endolymph, which is unique among extracellular fluids of the body, including the perilymph, in that its potassium ion concentration is higher (about 140 milliequivalents per litre) than its sodium ion concentration (about 15 milliequivalents per litre).

The physiology of hearing

Hearing is the process by which the ear transforms sound vibrations in the external environment into nerve impulses that are conveyed to the brain, where they are interpreted as sounds. Sounds are produced when vibrating objects, such as the plucked string of a guitar, produce pressure pulses of vibrating air molecules, better known as sound waves. The ear can distinguish different subjective aspects of a sound, such as its loudness and pitch, by detecting and analyzing different physical characteristics of the waves. Pitch is the perception of the frequency of sound waves—*i.e.,* the number of wavelengths that pass a fixed point in a unit of time. Frequency is usually measured in cycles per second, or hertz. The human ear is most sensitive to and most easily detects frequencies of 1,000 to 4,000 hertz, but at least for normal young ears the entire audible range of sounds extends from about 20 to 20,000 hertz. Sound waves of still higher frequency are referred to as ultrasonic, although they can be heard by other mammals. Loudness is the perception of the intensity of sound—*i.e.,* the pressure exerted by sound waves on the tympanic membrane. The greater their amplitude or strength, the greater is the pressure or intensity, and consequently the loudness, of the sound. The intensity of sound is measured and reported in decibels (dB), a unit that expresses the relative magnitude of a sound on a logarithmic scale. Stated in another way, the decibel is a unit for comparing the intensity of any given sound with a standard sound that is just perceptible to the normal human ear at a frequency in the range to which the ear is most sensitive. On the decibel scale, the range of human hearing extends from 0 dB, which represents a level that is all but inaudible, to about 130 dB, the level at which sound becomes painful. (See the article SOUND for a more in-depth discussion.)

Loudness and pitch

In order for a sound to be transmitted to the central nervous system, the energy of the sound undergoes three transformations (Figure 58). First, the air vibrations are converted to vibrations of the tympanic membrane and ossicles of the middle ear. These, in turn, become vibrations in the fluid within the cochlea. Finally, the fluid vibrations set up traveling waves along the basilar membrane

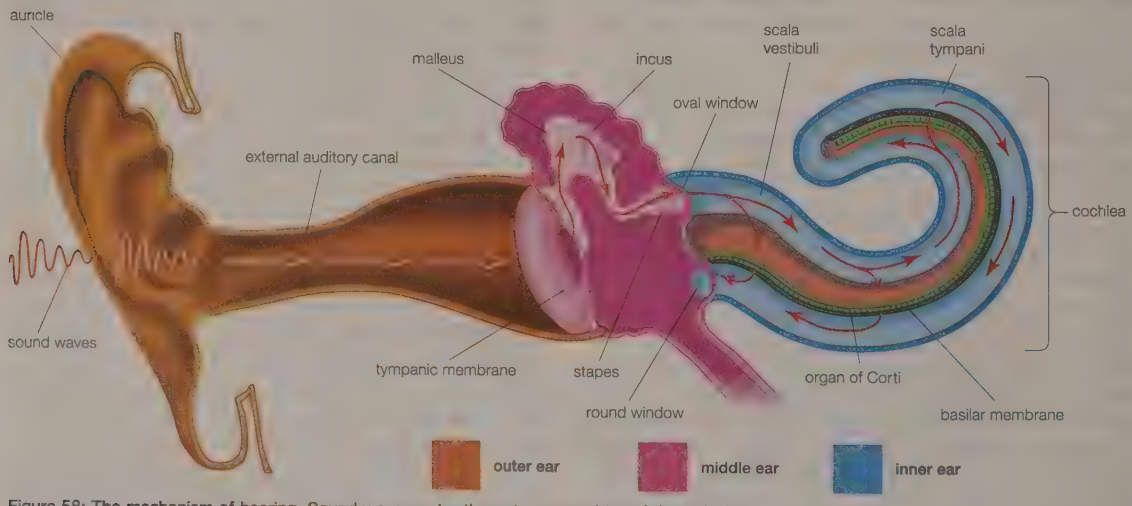


Figure 58: The mechanism of hearing. Sound waves enter the outer ear and travel through the external auditory canal until they reach the tympanic membrane, causing the membrane and the attached chain of auditory ossicles to vibrate. The motion of the stapes against the oval window sets up waves in the fluids of the cochlea, causing the basilar membrane to vibrate. This stimulates the sensory cells of the organ of Corti, atop the basilar membrane, to send nerve impulses to the brain.

Encyclopædia Britannica, Inc

that stimulate the hair cells of the organ of Corti. These cells convert the sound vibrations to nerve impulses in the fibres of the cochlear nerve, which transmits them to the brain stem, from which they are relayed, after extensive processing, to the primary auditory area of the cerebral cortex, the ultimate centre of the brain for hearing. Only when the nerve impulses reach this area does the listener become aware of the sound.

TRANSMISSION OF SOUND WAVES THROUGH THE OUTER AND MIDDLE EAR

Transmission of sound by air conduction. The outer ear directs sound waves from the external environment to the tympanic membrane (Figure 52). The auricle, the visible portion of the outer ear, collects sound waves and, with the concha, the cavity at the entrance to the external auditory canal, helps to funnel sound into the canal. Because of its small size and virtual immobility, the auricle in humans is less useful in sound gathering and direction finding than it is in many animals. The canal helps to enhance the amount of sound that reaches the tympanic membrane. This resonance enhancement works only for sounds of relatively short wavelength—those in the frequency range between 2,000 and 7,000 hertz—which helps to determine the frequencies to which the ear is most sensitive, those important for distinguishing the sounds of consonants.

Sounds reaching the tympanic membrane are in part reflected and in part absorbed. Only absorbed sound sets the membrane in motion. The tendency of the ear to oppose the passage of sound is called acoustic impedance (see below). The magnitude of the impedance depends on the mass and stiffness of the membrane and the ossicular chain and on the frictional resistance they offer.

When the tympanic membrane absorbs sound waves, its central portion, the umbo, vibrates as a stiff cone, bending inward and outward. The greater the force of the sound waves, the greater the deflection of the membrane and the louder the sound. The higher the frequency of a sound, the faster the membrane vibrates and the higher the pitch of the sound is. The motion of the membrane is transferred to the handle of the malleus, the tip of which is attached at the umbo. At higher frequencies the motion of the membrane is no longer simple, and transmission to the malleus may be somewhat less effective.

The malleus and incus are suspended by small elastic ligaments and are finely balanced, with their masses evenly distributed above and below their common axis of rotation. The head of the malleus and the body of the incus are tightly bound together, with the result that they move as a unit in unison with the tympanic membrane. At moderate sound pressures, the vibrations are passed on to the stapes, and the whole ossicular chain moves as a single mass. However, there may be considerable freedom of motion and some loss of energy at the joint between the incus and the stapes because of their relatively loose coupling. The stapes does not move in and out but rocks back and forth about the lower pole of its footplate, which impinges on the membrane covering the oval window in the bony plate of the inner ear. The action of the stapes transmits the sound waves to the perilymph of the vestibule and the scala vestibuli (Figure 58).

Function of the ossicular chain. In order for sound to be transmitted to the inner ear, the vibrations in the air must be changed to vibrations in the cochlear fluids. There is a challenge involved in this task that has to do with difference in impedance—the resistance to the passage of sound—between air and fluid. This difference, or mismatch, of impedances reduces the transmission of sound. The tympanic membrane and the ossicles function to overcome the mismatch of impedances between air and the cochlear fluids, and thus the middle ear serves as a transformer, or impedance matching device.

Matching of impedances

Ordinarily, when airborne sound strikes the surface of a body of water, almost all of its energy is reflected and only about 0.1 percent passes into the water. In the ear this would represent a transmission loss of 30 decibels, enough to seriously limit the ear's performance, were it not for the transformer action of the middle ear. The matching of impedances is accomplished in two ways, primarily by the reduction in area between the tympanic membrane and the stapes footplate and secondarily by the mechanical advantage of the lever formed by the malleus and incus. Although the total area of the tympanic membrane is about 69 square millimetres (0.1 square inch), the area of its central portion that is free to move has been estimated at about 43 square millimetres. The sound energy that causes this area of the membrane to vibrate is transmitted and concentrated in the 3.2-square-millimetre area of the stapes footplate. Thus, the pressure is increased at least 13 times. The mechanical advantage of the ossicular lever (which exists because the handle of the malleus is longer than the long projection of the incus) amounts to about 1.3. The total increase in pressure at the footplate is, therefore, not less than 17-fold, depending on the area of the tympanic membrane that is actually vibrating. At frequencies in the range of 3,000 to 5,000 hertz, the increase may be even greater because of the resonant properties of the ear canal.

The ossicular chain not only concentrates sound in a small area but also applies sound preferentially to one window of the cochlea, the oval window (Figure 58). If the oval and round windows were exposed equally to airborne sound crossing the middle ear, the vibrations in the perilymph of the scala vestibuli would be opposed by those in the perilymph of the scala tympani, and little effective movement of the basilar membrane would result. As it is, sound is delivered selectively to the oval window, and the round window moves in reciprocal fashion, bulging outward in response to an inward movement of the stapes footplate and inward when the stapes moves away from the oval window. The passage of vibrations through the air across the middle ear from the tympanic membrane to the round window is of negligible importance.

Function of the muscles of the middle ear. The muscles of the middle ear, the tensor tympani and the stapedius, can influence the transmission of sound by the ossicular chain. Contraction of the tensor tympani pulls the handle of the malleus inward and, as the name of the muscle suggests, tenses the tympanic membrane. Contraction of the stapedius pulls the stapes footplate outward from the oval window and thereby reduces the intensity of sound reaching the cochlea. The stapedius responds reflexly with quick contraction to sounds of high intensity applied either to the same ear or to the opposite ear. The reflex has been likened to the blink of the eye or the constriction of the pupil of the eye in response to light and is thought to have protective value. Unfortunately, the contractions of the middle-ear muscles are not instantaneous, so that they do not protect the cochlea against damage by sudden intense noise, such as that of an explosion or of gunfire. They also fatigue rather quickly and thus offer little protection against injury sustained from high-level noise, such as that experienced in rock concerts and many industrial workplaces.

Transmission of sound by bone conduction. There is another route by which sound can reach the inner ear: by conduction through the bones of the skull. When the handle of a vibrating tuning fork is placed on a bony prominence such as the forehead or mastoid process behind the ear, its note is clearly audible. Similarly, the ticking of a watch held between the teeth can be distinctly heard. When the external canals are closed with the fingers, the sound becomes louder, indicating that it is not entering the ear by the usual channel. Instead, it is producing vibrations of the skull that are passed on to the inner ear, either directly or indirectly, through the bone.

The higher audible frequencies cause the skull to vibrate in segments, and these vibrations are transmitted to the cochlear fluids by direct compression of the otic capsule, the bony case enclosing the inner ear. Because the round window membrane is more freely mobile than the stapes footplate, the vibrations set up in the perilymph of the scala vestibuli are not canceled out by those in the scala tympani, and the resultant movements of the basilar membrane can stimulate the organ of Corti. This type of transmission is known as compression bone conduction.

Compression and inertial bone conduction

At lower frequencies—*i.e.*, 1,500 hertz and below—the skull moves as a rigid body. The ossicles are less affected

and move less freely than the cochlea and the margins of the oval window because of their inertia, their suspension in the middle-ear cavity, and their loose coupling to the skull. The result is that the oval window moves with respect to the footplate of the stapes, which gives the same effect as if the stapes itself were vibrating. This form of transmission is known as inertial bone conduction.

TRANSMISSION OF SOUND WITHIN THE INNER EAR

Transmission of sound waves in the cochlea. The mechanical vibrations of the stapes footplate at the oval window creates pressure waves in the perilymph of the scala vestibuli of the cochlea. These waves move around the tip of the cochlea through the helicotrema into the scala tympani and dissipate as they hit the round window (Figure 58). The wave motion is transmitted to the endolymph inside the cochlear duct. As a result the basilar membrane vibrates, which causes the organ of Corti to move against the tectorial membrane, stimulating generation of nerve impulses to the brain.

Analysis of sound

Within the cochlea the different frequencies of complex sounds are sorted out, or analyzed, and the physical energy of these sound vibrations is converted, or transduced, into electrical impulses that are transmitted to the brain stem by the cochlear nerve. The cochlea analyzes sound frequencies (distinguishes pitch) by means of the basilar membrane, which exhibits different degrees of stiffness, or resonance, along its length (Figure 59).

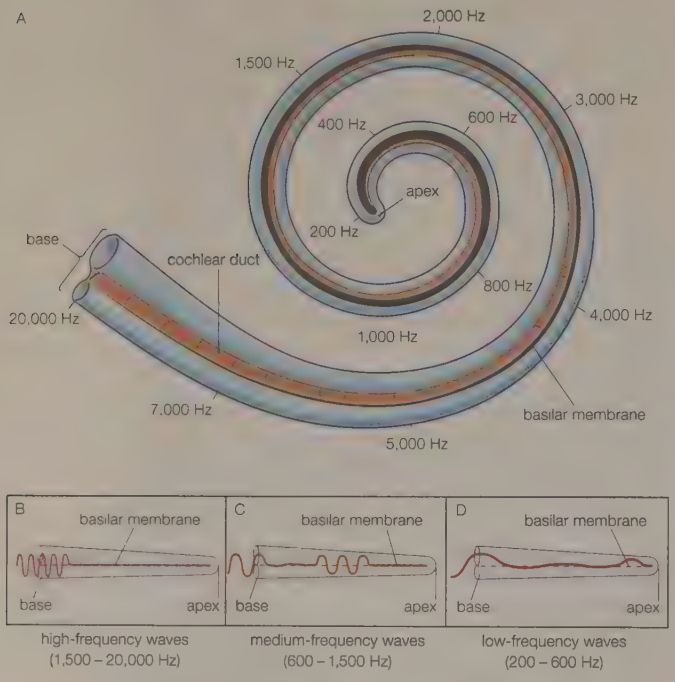


Figure 59: The analysis of sound frequencies by the basilar membrane. (A) The fibres of the basilar membrane become progressively wider and more flexible from the base of the cochlea to the apex. As a result, each area of the basilar membrane vibrates preferentially to a particular sound frequency. (B) High-frequency sound waves cause maximum vibration of the area of the basilar membrane nearest to the base of the cochlea, (C) medium-frequency waves affect the centre of the membrane, (D) and low-frequency waves preferentially stimulate the apex of the basilar membrane. (The locations of cochlear frequencies along the basilar membrane shown are a composite drawn from different sources.)
Encyclopædia Britannica, Inc

Pitch is distinguished because of the continuous changes that occur along the length of the basilar membrane, which increases in width and mass and decreases in stiffness from its base near the oval window to its apex. Each region of the membrane is most affected by a specific frequency of vibrations. Low-frequency sounds cause the apical end of the membrane to vibrate, and high-frequency sounds cause the basal end to vibrate (Figure 59). Vibrations reaching the basal end through the perilymph proceed along the membrane as traveling waves that attain their maximum amplitude at a distance corresponding to their frequency and then rapidly subside. The higher the frequency of the sound imposed, the shorter the distance the waves travel. Thus, a tone of a given frequency causes stimulation to reach a peak at a certain place on the basilar membrane. The region that vibrates most vigorously stimulates the greatest number of hair cells in that area of the organ of Corti, and these hair cells send the most nerve impulses to the auditory nerve and the brain (see below). The brain recognizes the place on the basilar membrane and thus the pitch of the tone by the particular group of nerve fibres activated. For the lower frequencies—up to about 3,000 hertz—the rate of stimulation is also an important indicator of pitch. This means that the auditory nerve fibres convey information to the brain about the timing of the sound frequency as well as its place of maximum vibration on the membrane. For higher frequencies place alone seems to be decisive.

Loudness also is determined at this level by the amplitude, or height, of the vibration of the basilar membrane. As a sound increases, so does the amplitude of the vibration. This increases both the number of hair cells stimulated and the rate at which they generate nerve impulses.

Transduction of mechanical vibrations. The hair cells located in the organ of Corti transduce mechanical sound vibrations into nerve impulses. They are stimulated when the basilar membrane, on which the organ of Corti rests, vibrates. The hair cells are held in place by the reticular lamina, a rigid structure supported by the pillar cells, or rods of Corti, which are attached to the basilar fibres. At the base of the hair cells is a network of cochlear nerve endings, which lead to the spiral ganglion of Corti in the modiolus of the cochlea. The spiral ganglion sends axons into the cochlear nerve. At the top of the hair cell is a hair bundle containing stereocilia, or sensory hairs, that project upward into the tectorial membrane, which lies above the stereocilia in the cochlear duct. (The single kinocilium, which is found on the hair cells of the vestibular system, is not found on the receptor cells of the cochlea.) When the basilar membrane moves upward, the reticular lamina moves upward and inward; when the membrane moves downward, the reticular lamina moves downward and outward. The resultant shearing forces between the reticular lamina and the tectorial membrane displace or bend the longest of the stereocilia, exciting the nerve fibres at the base of the hair cells.

The mechanism the hair cell uses to convert sound into an electrical stimulus is not completely understood, but certain key features are known. One of the most important aspects of this process is the endocochlear potential, which exists between the endolymph and perilymph. This direct current potential difference is about +80 millivolts and results from the difference in potassium content between the two fluids. It is thought to be maintained by the continual transport of potassium ions from the perilymph into the cochlear duct by the stria vascularis. The endolymph, which has a high potassium level and a positive potential, is contained in the cochlear duct and thus bathes the tops of the hair cells. The perilymph, which has a low potassium level and a negative potential, is contained in the scala vestibuli and scala tympani and bathes the lower parts of the hair cells. The inside of the hair cell has a negative intracellular potential of −60 millivolts with respect to the perilymph and −140 millivolts with respect to the endolymph. This rather steep gradient, especially at the tip of the cell, is thought to sensitize the cell to the slightest sound.

Endocochlear potential

The stereocilia are graded in height, becoming longer on the side away from the modiolus. All the stereocilia are interlinked so that, when the taller ones are moved against the tectorial membrane, the shorter ones move as well. The mechanical movement of this hair bundle generates an alternating hair cell receptor potential. This occurs in the following manner. When the stereocilia are bent in the direction of increasing stereocilia length, ion channels in the membrane open, allowing potassium ions to move into the cell. The influx of potassium ions excites, or depolarizes, the hair cell. However, when the stereocilia

are deflected in the opposite direction, the ion channels are shut and the hair cell is inhibited, or hyperpolarized. The depolarization of the cell stimulates the release of chemicals called neurotransmitters from the base of the hair cell. The neurotransmitters are absorbed by the nerve fibres located at the basal end of the hair cell, stimulating them to send an electrical signal along the cochlear nerve.

COCHLEAR NERVE AND CENTRAL AUDITORY PATHWAYS

Auditory nerve fibres. The vestibulocochlear nerve consists of two anatomically and functionally distinct parts: the cochlear nerve, which innervates the organ of hearing, and the vestibular nerve, which innervates the organs of equilibrium. The fibres of the cochlear nerve originate from an aggregation of nerve cell bodies, the spiral ganglion, located in the modiolus of the cochlea. The neurons of the spiral ganglion are called bipolar cells because they have two sets of processes, or fibres, that extend from opposite ends of the cell body. The longer, central fibres, also called the primary auditory fibres, form the cochlear nerve, and the shorter, peripheral fibres extend to the bases of the inner and outer hair cells. They extend radially from the spiral ganglion to the habenula perforata, a series of tiny holes beneath the inner hair cells. At this point they lose their myelin sheaths and enter the organ of Corti as thin, unmyelinated fibres. There are only about 30,000 of these fibres, and the greater number of them—about 95 percent—innervate the inner hair cells. The remainder cross the tunnel of Corti to innervate the outer hair cells. The longer central processes of the bipolar cochlear neurons unite and are twisted like the cords of a rope to form the cochlear nerve trunk. These primary auditory fibres exit the modiolus through the internal meatus, or passageway, and immediately enter the part of the brain stem called the medulla.

Auditory pathways. *Ascending pathways.* The central auditory pathways extend from the medulla to the cerebral cortex. They consist of a series of nuclei (groups of nerve cell bodies in the central nervous system similar to a peripheral ganglion) connected by fibre tracts made up of their axons (processes that convey signals away from the cell bodies). This complex chain of nerve cells helps to process and relay auditory information, encoded in the form of nerve impulses, directly to the highest cerebral levels in the cortex of the brain. To some extent different properties of the auditory stimulus are conveyed along distinct parallel pathways. This method of transmission, employed by other sensory systems, provides a means for the central nervous system to analyze different properties of the single auditory stimulus, with some information processed at low levels and other information at higher levels. At lower levels of the pathway, information as to pitch, loudness, and localization of sounds is processed, and appropriate responses, such as the contraction of the intra-aural muscles, turning of the eyes and head, or movements of the body as a whole, are initiated.

Processing of auditory information

In the medulla the fibres of the cochlear nerve terminate when they reach a collection of nerve cells called the cochlear nucleus. The cochlear nucleus consists of several distinct cell types and is divided into the dorsal and ventral cochlear nucleus. Each cochlear nerve fibre branches at the cochlear nucleus, sending one branch to the dorsal and the other branch to the ventral cochlear nucleus.

Some fibres from the ventral cochlear nucleus pass across the midline to the cells of the superior olivary complex, whereas others make connection with the olivary cells of the same side. Together, these fibres form the trapezoid body. Fibres from the dorsal cochlear nucleus cross the midline to end on the cells of the nuclei of the lateral lemniscus. There they are joined by the fibres from the ventral cochlear nuclei of both sides and from the olivary complex. The lemniscus is a major tract, most of the fibres of which end in the inferior colliculus, the auditory centre of the midbrain, although some fibres may bypass the colliculus and end, together with the fibres from the colliculus, at the next higher level, the medial geniculate body. From the medial geniculate body there is an orderly projection of fibres to a portion of the cortex of the temporal lobe.

In humans and other primates the primary acoustic area in the cerebral cortex is the superior transverse temporal gyri of Heschl, a ridge in the temporal lobe, on the lower lip of the deep cleft between the temporal and parietal lobes, known as the sylvian fissure.

Because about half of the fibres of the auditory pathways cross the midline while others ascend on the same side of the brain, each ear is represented in both the right and left cortex. For this reason, even when the auditory cortical area of one side is injured by trauma or stroke, binaural hearing may be little affected. Impaired hearing due to bilateral cortical injury involving both auditory areas has been reported, but it is extremely rare.

Descending pathways. Parallel with the pathway ascending from the cochlear nuclei to the cortex is a pathway descending from the cortex to the cochlear nuclei. In both pathways some of the fibres remain on the same side, while others cross the midline to the opposite side of the brain. There is also evidence of a "spur" line ascending from the dorsal cochlear nucleus to the cerebellum and another descending from the inferior colliculus to the cerebellum. The significance of these cerebral connections is not clear, but they may antedate the evolutionary development of the cerebral cortex. In general, the descending fibres may be regarded as exercising an inhibitory function by means of a sort of "negative feedback." They also may determine which ascending impulses are to be blocked and which are allowed to pass on to the higher centres of the brain.

Function of descending fibres

From the superior olivary complex, a region in the medulla oblongata, there arises also a fibre tract called the olivocochlear bundle. It constitutes an efferent system, or feedback loop, by which nerve impulses, thought to be inhibitory, reach the hair cells. This system, which uses acetylcholine as a neurotransmitter, is presumably involved in sharpening, or otherwise modifying, the analysis that is made in the cochlea.

Analysis of sound by the auditory nervous system. Evidence of orderly spatial representations of the organ of Corti at the lower levels of the auditory pathway has been reported by many investigators. These patterns seem to be in accord with the place theory of the cochlear analysis of sound. Physiological evidence of tuning of the auditory system also has been obtained by recording with the electrical potentials from individual neurons at various levels. Most neurons of the auditory pathway show a "best frequency"—*i.e.,* a frequency to which the individual neuron responds at minimal intensity. This finding is entirely compatible with experimental evidence of frequency tuning of the hair cells (see above *Transmission of sound within the inner ear*). With each increase in the intensity of the sound stimulus, the neuron is able to respond to a wider band of frequencies, thus reflecting the broad tuning of the basilar membrane. With sounds of lower frequency, the rate of impulses fired by the neuron reflects the stimulus frequency, and the response often reveals phase-locking with the stimulus; that is, the nerve fibres are stimulated at regularly recurring intervals, corresponding to a particular position or phase, of each sound wave. Increased intensity of stimulation causes a more rapid rate of responding. In general, the pitch of a sound tends to be coded in terms of which neurons are responding, and its loudness is determined by the rate of response and the total number of neurons activated.

It appears likely that in humans the cortex is not involved in frequency recognition but is reserved for the analysis of more complex auditory stimuli, such as speech and music, for which the temporal sequence of sounds is equally important. Presumably it is also at the cortical level that the meaning of sounds is interpreted and behaviour is adjusted in accordance with their significance. Such functions were formerly attributed to an "auditory association area" immediately surrounding the primary area, but they probably should be thought of as involving much more of the cerebral cortex, thanks to the multiple, parallel interconnections between the various areas.

The localization of sounds from a stationary source in the horizontal plane is known to depend on the recognition of minute differences in the intensity and time of arrival of the sound at the two ears. A sound that arrives

Localization of sound

at the right ear a few microseconds sooner than it does at the left or that sounds a few decibels louder in that ear is recognized as coming from the right. In a real-life situation the head may also be turned to pinpoint the sound by facing it and thus canceling these differences. For low-frequency tones a difference in phase at the two ears is the criterion for localization, but for higher frequencies the difference in loudness caused by the sound shadow of the head becomes all-important. Such comparisons and discriminations appear to be carried out at brain stem and midbrain levels of the central auditory pathway. The spectral shapes of sounds have been shown to be most important for determining the elevation of a source that is not in the horizontal plane. Localization of sound that emanates from a moving source is a more complicated task for the nervous system and apparently involves the cerebral cortex and short-term memory. Experiments in animals have shown that injury to the auditory area of the cortex on one side of the brain interferes with the localization of a moving sound source on the opposite side of the body.

Each cochlear nucleus receives impulses only from the ear of the same side. A comparison between the responses of the two ears first becomes possible at the superior olivary complex, which receives fibres from both cochlear nuclei. Electrophysiological experiments in animals have shown that some neurons of the accessory nucleus of the olivary complex respond to impulses from both ears. Others respond to impulses from one side exclusively, but their response is modified by the simultaneous arrival of impulses from the other side.

The system appears to be capable of making the extraordinarily fine discriminations of time and intensity that are necessary for sound localization. By virtue of such bilateral neural interconnections in the brain, the two ears together can be much more effective than one ear alone in picking out a particular sound in the presence of a background of noise. They also permit attention to be directed to a single source of sound, such as one instrument in an orchestra or one voice in a crowd.

HEARING TESTS

Before the development of electroacoustic equipment for generating and measuring sound, the available tests of hearing gave approximate answers at best. A person's hearing could be specified in terms of the ability to distinguish the ticking of a watch or the clicking of coins or the distance at which conversational speech or a whispered voice could be understood. The examiner also might note the length of time the person could hear the gradually diminishing note of a tuning fork, comparing the performance with his own.

Tuning fork tests. A qualitative assessment of hearing loss can be carried out with a tuning fork. These tests exploit the ability of sound to be conducted through the bones of the skull (see above *Transmission of sound waves through the outer and middle ear: Transmission of sound by bone conduction*).

One example of a tuning fork test is the Weber test, in which the fork is simply placed on the person's forehead and the examiner asks in which ear the person hears it. If a sensorineural lesion is present in one ear, the person will localize the sound in the opposite, or "better," ear. If a conductive defect is present, the person will localize it in the "worse" ear—*i.e.,* the one that is protected from interference by extraneous sounds. This simple test has been a valuable aid in the diagnosis of otosclerosis for many years.

Audiometry. With the introduction of the electric audiometer in the 1930s, it became possible to measure an individual's hearing threshold for a series of pure tones ranging from a lower frequency of 125 hertz to an upper frequency of 8,000 or 10,000 hertz. This span includes the three octaves between 500 and 4,000 hertz that are most important for speech.

The audiometer consists of an oscillator or signal generator, an amplifier, a device called an attenuator, which controls and specifies the intensity of tones produced, and an earphone or loudspeaker. The intensity range is usually 100 decibels in steps of 5 decibels. The "zero dB" level represents normal hearing for young adults under favourable, noise-free laboratory conditions. It was established in 1964 as an international standard.

Pure-tone audiometry

In pure-tone audiometry each ear is tested separately, while the other is shielded against sound. The person being tested wears an earphone or sits in front of a loudspeaker in a quiet test chamber, with instructions to give a hand signal whenever a brief tone is sounded. The audiologist proceeds to determine the lowest intensity for each frequency at which the person reports being just able to hear the tone 50 percent of the time. For example, one who hears the tone of 4,000 hertz only half the time at the 40-decibel setting has a 40-decibel hearing level for that frequency—*i.e.,* a threshold 40 decibels above the normal threshold. A graph showing the hearing level for each ear by octaves and half octaves across the frequency range of 125 to 8,000 hertz is called an audiogram. The shape of the audiogram for an individual who is hard-of-hearing can provide the otologist or audiologist with important information for determining the nature and cause of the hearing defect.

A calibrated bone-conduction vibrator usually is furnished with the audiometer so that hearing by bone conduction also can be measured. When an individual has otosclerosis or another conductive defect of the middle ear, there may be a sizable difference between the air-conduction and bone-conduction audiograms, the so-called air-bone gap. This difference is a measure of the loss in transmission across the middle ear and indicates the maximum improvement that may be obtained through successful corrective surgery. When the defect is confined to the organ of Corti, the bone-conduction audiogram shows the same degree of loss as the air-conduction audiogram. In such cases of sensorineural impairment, surgery is seldom capable of improving hearing, but a hearing aid may be helpful.

Although faint sounds may not be heard at all by the ear with a sensorineural impairment, more intense sounds may be as loud as they are to a healthy ear. This rapid increase in loudness above the threshold level is called recruitment. When the opposite ear has normal hearing, recruitment can be measured by the alternate binaural loudness balance test. The subject is asked to set the controls so that the loudness of the tone heard in the defective ear matches that of the tone heard in the normal ear. By repeating the comparison at several intensity levels, the presence or absence of recruitment can be demonstrated. When recruitment is excessive, the range of useful hearing between the threshold and the level at which loudness becomes uncomfortable or intolerable may be narrow, so that the amplification provided by a hearing aid is of limited value to the subject.

Recruitment

Although hearing thresholds for pure tones give some indication of the person's ability to hear speech, direct measurement of this ability is also important. Two types of tests are used most often. In one test the speech reception threshold is measured by presenting words of spondee pattern—*i.e.,* words containing two syllables of equal emphasis, as in "railway" or "football"—at various intensity levels until the level is found at which the person can just hear and repeat half the words correctly. This level usually corresponds closely to the average of the person's thresholds for frequencies of 500, 1,000, and 2,000 hertz. A more important measure of socially useful hearing is the discrimination score. For this test a list of selected monosyllabic words is presented at a comfortable intensity level, and the subject is scored in terms of the percentage of words heard correctly. This test is helpful in evaluating certain forms of hearing impairment in which the sounds may be audible but words remain unintelligible. Such tests usually are carried out in a quiet, sound-treated room that excludes extraneous noise. These tests may give an overly optimistic impression of the ability of the individual with a sensorineural impairment to understand speech in ordinary noisy surroundings. For this reason speech tests are best carried out against a standardized noise background as well as in the quiet. A person with a conductive defect may be less disturbed by the noisy environment than a

healthy subject. More elaborate tests, which often involve speech or sound localization, are available for evaluating hearing when central defects of the auditory system are suspected as a result of aging, disease, or injury. Their interpretation may be difficult, however, and the diagnostic information they furnish may be unclear.

When the hearing of infants or others who are unable to cooperate in standard audiometric tests must be measured, their thresholds for pure tones can be established by electrophysiological means. One type of test is the electrocochleogram (ECoG). Electric potentials representing impulses in the cochlear nerve are recorded from the outer surface of the cochlea by means of a fine, insulated needle electrode inserted through the tympanic membrane to make contact with the promontory of the basal turn. This test provides a direct sampling of cochlear function.

A noninvasive, painless, and more frequently used test is brain-stem-evoked response audiometry (BERA). In this test electrodes are pasted to the skin (one placed behind the ear) and are used to record the neural responses to brief tones. The minute potentials evoked by a train of brief sound stimuli are suitably amplified and averaged by a small computer to cancel out background activity, such as potentials from muscles or the cerebral cortex. The typical recording shows a series of five or six waves that represent the responses of successive neural centres of the auditory pathway of the brain stem and provide information about the strength and timing of their activity.

The physiology of balance: vestibular function

Organs of balance

The vestibular system is the sensory apparatus of the inner ear that helps the body maintain its postural equilibrium. The information furnished by the vestibular system is also essential for coordinating the position of the head and the movement of the eyes. There are two sets of end organs in the inner ear, or labyrinth: the semicircular canals, which respond to rotational movements (angular acceleration); and the utricle and saccule within the vestibule, which respond to changes in the position of the head with respect to gravity (linear acceleration; see Figure 60). The information these organs deliver is proprioceptive in character, dealing with events within the body itself, rather than exteroceptive, dealing with events outside the body, as in the case of the responses of the cochlea to sound. Functionally these organs are closely related to the cerebellum and to the reflex centres of the spinal cord and brain stem that govern the movements of the eyes, neck, and limbs. For anatomical descriptions of the vestibular apparatus see above *Anatomy of the human ear: Inner ear: Vestibular system*.

Although the vestibular organs and the cochlea are derived embryologically from the same formation, the otic vesicle, their association in the inner ear seems to be a matter more of convenience than of necessity. From both the developmental and the structural point of view, the kinship of the vestibular organs with the lateral line system of the fish is readily apparent. The lateral line system is made up of a series of small sense organs located in the skin of the head and along the sides of the body of fishes. Each organ contains a crista, sensory hair cells, and a cupula, as found in the ampullae of the semicircular ducts. The cristae respond to waterborne vibrations and to pressure changes.

The anatomists of the 17th and 18th centuries assumed that the entire inner ear, including the vestibular apparatus, is devoted to hearing. They were impressed by the orientation of the semicircular canals, which lie in three planes more or less perpendicular to one another, and believed that the canals must be designed for localizing a source of sound in space. The first investigator to present evidence that the vestibular labyrinth is the organ of equilibrium was a French experimental neurologist, Marie-Jean-Pierre Flourens, who in 1824 reported a series of experiments in which he had observed abnormal head movements in pigeons after he had cut each of the semicircular canals in turn. The plane of the movements was always the same as that of the injured canal. Hearing was not affected when he cut the nerve fibres to these organs, but it was abolished when he cut those to the basilar papilla (the bird's uncoiled cochlea). It was not until almost half a century later that the significance of his findings was appreciated and the semicircular canals were recognized as sense organs specifically concerned with the movements and position of the head.

DETECTION OF ANGULAR ACCELERATION: DYNAMIC EQUILIBRIUM

Because the three semicircular canals—superior, posterior, and horizontal—are positioned at right angles to one another, they are able to detect movements in three-dimensional space (see above *Anatomy of the human ear: Inner ear: Semicircular canals*). When the head begins to rotate in any direction, the inertia of the endolymph causes it

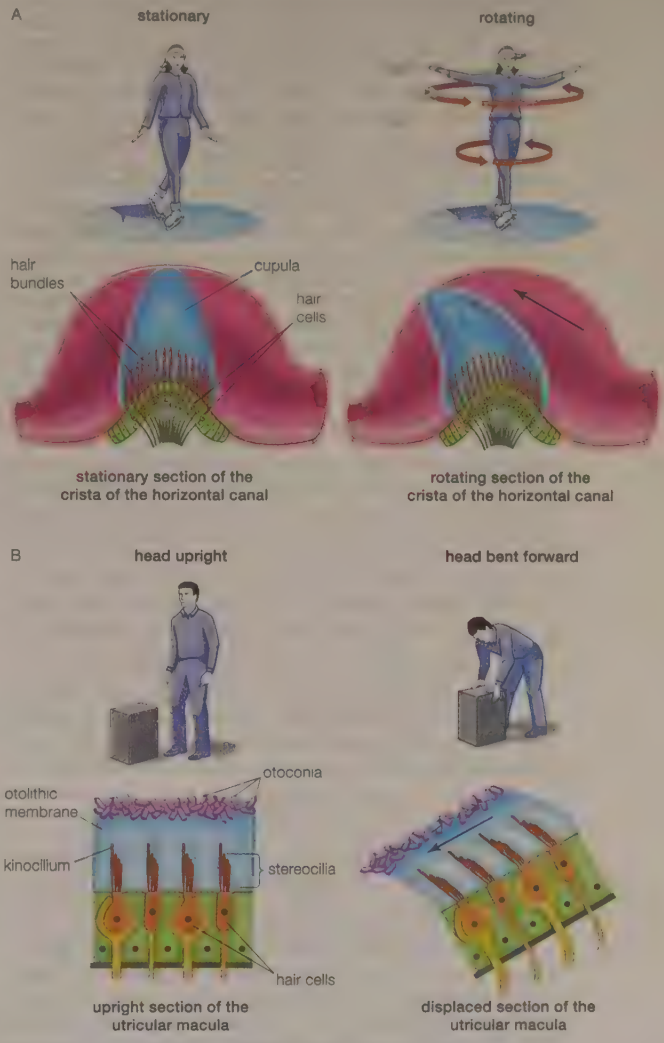


Figure 60: The physiology of balance. Each of the two sensory organs of balance, the cristae of the semicircular ducts and the maculae of the utricle and saccule, evaluates different aspects of equilibrium, but both function in a similar way. (A) The cristae respond to rotational movements and are involved in dynamic equilibrium. (B) The maculae respond to gravitational pull and help to maintain static equilibrium.

Encyclopædia Britannica, Inc

to lag behind, exerting pressure that deflects the cupula in the opposite direction. This deflection stimulates the hair cells by bending their stereocilia in the opposite direction (Figure 60A). The German physiologist Friedrich Goltz formulated the "hydrostatic concept" in 1870 to explain the working of the semicircular canals. He postulated that the canals are stimulated by the weight of the fluid they contain, the pressure it exerts varying with the head position. In 1873 the Austrian scientists Ernst Mach and Josef Breuer and the Scottish chemist Crum Brown, working independently, proposed the "hydrody-

Discovery of the function of the semicircular canals

namic concept," which held that head movements cause a flow of endolymph in the canals and that the canals are then stimulated by the fluid movements or pressure changes. The German physiologist J.R. Ewald showed that the compression of the horizontal canal in a pigeon by a small pneumatic hammer causes endolymph movement toward the crista and turning of the head and eyes toward the opposite side. Decompression reverses both the direction of endolymph movement and the turning of the head and eyes. The hydrodynamic concept was proved correct by later investigators who followed the path of a droplet of oil that was injected into the semicircular canal of a live fish. At the start of rotation in the plane of the canal the cupula was deflected in the direction opposite to that of the movement and then returned slowly to its resting position. At the end of rotation it was deflected again, this time in the same direction as the rotation, and then returned once more to its upright stationary position. These deflections resulted from the inertia of the endolymph, which lags behind at the start of rotation and continues its motion after the head has ceased to rotate. The slow return is a function of the elasticity of the cupula itself.

These opposing deflections of the cupula affect the vestibular nerve in different ways, which have been demonstrated in experiments involving the labyrinth removed from a cartilaginous fish. The labyrinth, which remained active for some time after its removal from the animal, was used to record vestibular nerve impulses arising from one of the ampullar cristae. When the labyrinth was at rest there was a slow, continuous, spontaneous discharge of nerve impulses, which was increased by rotation in one direction and decreased by rotation in the other. In other words, the level of excitation rose or fell depending on the direction of rotation.

The deflection of the cupula excites the hair cells by bending the cilia atop them: deflection in one direction depolarizes the cells; deflection in the other direction hyperpolarizes them. Electron-microscopic studies have shown how this polarization occurs. The hair bundles in the cristae are oriented along the axis of each canal. For example, each hair cell of the horizontal canals has its kinocilium facing toward the utricle, whereas each hair cell of the superior canals has its kinocilium facing away from the utricle. In the horizontal canals deflection of the cupula toward the utricle—*i.e.*, bending of the stereocilia toward the kinocilium—depolarizes the hair cells and increases the rate of discharge. Deflection away from the utricle causes hyperpolarization and decreases the rate of discharge. In superior canals these effects are reversed.

DETECTION OF LINEAR ACCELERATION: STATIC EQUILIBRIUM

The gravity receptors that respond to linear acceleration of the head are the maculae of the utricle and saccule (see above *Anatomy of the human ear: Inner ear: Vestibule*). The left and right utricular maculae are in the same, approximately horizontal, plane and because of this position are more useful in providing information about the position of the head and its side-to-side tilts when a person is in an upright position. The saccular maculae are in parallel vertical planes and probably respond more to forward and backward tilts of the head.

Stimulation of the maculae

Both pairs of maculae are stimulated by shearing forces between the otolithic membrane and the cilia of the hair cells beneath it (Figure 60B). The otolithic membrane is covered with a mass of minute crystals of calcite (otoconia), which add to the membrane's weight and increase the shearing forces set up in response to a slight displacement when the head is tilted. The hair bundles of the macular hair cells are arranged in a particular pattern—facing toward (in the utricle) or away from (in the saccule) a curving midline—that allows detection of all possible head positions. These sensory organs, particularly the utricle, have an important role in the righting reflexes and in reflex control of the muscles of the legs, trunk, and neck that keep the body in an upright position. The role of the saccule is less completely understood. Some investigators have suggested that it is responsive to vibration as well as to linear acceleration of the head in the sagittal (fore and aft) plane. Of the two receptors, the utricle appears to be the dominant partner. There is evidence that the mammalian saccule may even retain traces of its sensitivity to sound inherited from the fishes, in which it is the organ of hearing. (J.E.H.)

Ear diseases and hearing disorders

Impaired hearing is, with rare exception, the result of disease or abnormality of the outer, middle, or inner ear. Serious impairment of hearing at birth almost always results from a dysfunction of the auditory nerve and cannot be improved by medical or surgical treatment. In early and late childhood the most frequent cause for impaired hearing is poor functioning of the eustachian tubes with the accumulation of a clear, pale yellowish fluid in the middle-ear cavity, a disorder called serous, or secretory, otitis media. In early and middle adult life the usual cause for progressive impairment of hearing is otosclerosis. The usual cause of hearing loss after the age of 60 is presbycusis, a disorder that results from the aging process.

The structure and function of the human auditory and vestibular systems are treated in detail above. This section deals with the more important diseases and disorders of the outer, middle, and inner ear.

OUTER EAR

Diseases of the outer ear are those that afflict skin, cartilage, and the glands and hair follicles in the outer-ear canal. The sound-transmitting function of the outer ear is impaired when the ear canal becomes filled with tumour, infected material, or earwax (cerumen), so that sound cannot reach the tympanic membrane, or eardrum. The most common diseases of the outer ear are briefly described in the following paragraphs.

Infections and injuries. *Frostbite.* The exposed position of the outer ear makes it the part of the body most frequently affected by freezing, or frostbite. Humidity, duration of exposure, and, most of all, wind, in addition to degrees of temperature below freezing, predispose to the occurrence of frostbite. The frozen area begins along the upper and outer edge of the ear, which becomes yellow-white and waxy in appearance, cold and hard to the touch, and numb with loss of skin sensation.

In treatment of frostbite the victim is placed as soon as possible in a warm room, but the frozen ear is kept cool by applying ice wrapped in a towel until the returning blood circulation gradually thaws the frozen part from within. Massage of the frozen ear is avoided, for it is likely to injure the skin. Heat applied to the frozen area before circulation is established can result in clotting of the blood in the blood vessels. This in turn can result in death of that part of the ear, which turns black and eventually falls off, a process called dry gangrene.

Hematoma. Injury to the outer ear can cause bleeding between the cartilage and the skin, producing a smooth, rounded, nontender purplish swelling called hematoma. The accumulation of clotted blood is removed by a surgeon because, if it is left, it will become transformed into scar tissue and cause a permanent, irregular thickening of the outer ear commonly called cauliflower ear and seen in boxers and wrestlers whose ears receive much abuse.

Infections of the outer ear

Perichondritis. Infection of the cartilage of the outer ear, called perichondritis, is unusual but may occur from injury or from swimming in polluted water. It is due to a particular microorganism, *Pseudomonas aeruginosa.* There is a greenish or brownish, musty or foul-smelling discharge from the outer-ear canal, while the affected outer ear becomes tender, dusky red, and two to three times its normal thickness. Prompt antibiotic treatment is necessary to prevent permanent deformity of the outer ear.

External otitis. Infection of the outer-ear canal by molds or various microorganisms occurs especially in warm, humid climates and among swimmers. The ear canal itches and becomes tender; a small amount of thin, often foul-smelling material drains from it. If the canal becomes clogged by the swelling and drainage, hearing will be impaired. Careful and thorough cleaning of the outer-ear canal by a physician, application of antiseptic

or antibiotic eardrops, and avoidance of swimming are indicated to clear up the infection.

Boil in the ear (furuncle). Infection of a hair follicle anywhere on the body is known as a boil, or furuncle. This can occur in a hair follicle in the outer-ear canal, especially when there is infection of the skin of the canal. It always occurs because of a particular type of germ known as staphylococcus. Because the skin of the ear canal is closely attached to the underlying cartilage, a boil in the ear canal is especially painful, with swelling, redness, and tenderness but generally without fever. Heat applied to the outer ear by a hot-water bottle or electric pad helps the infection to come to a head and begin to drain. Treatment with a systemic and local antibiotic is required to prevent other hair follicles from becoming infected.

Erysipelas of the outer ear. Erysipelas is an infection in the skin caused by a particular type of streptococcus and characterized by a slowly advancing red, slightly tender thickening of the skin. It may begin at the ear and spread to the face and neck. Centuries ago erysipelas epidemics caused severe and often fatal infections. In AD 1089 one of the most severe erysipelas epidemics occurred. The disease was referred to as St. Anthony's fire because those who prayed to St. Anthony were said to recover; others, who did not, died. Today erysipelas is usually a mild and comparatively rare infection that clears up rapidly when treated with an antibiotic.

Osteoma of the bony ear canal. Osteoma of the bony ear canal is a bony knob that grows close to the tympanic membrane, especially in those who swim a great deal in cold water. It is not dangerous and does not need to be removed unless the bony overgrowth becomes large enough to block the ear canal.

Cyst of the ear. A cyst is a sac filled with liquid or semisolid material. A cyst of the ear is most often caused by a gland that lubricates the skin behind the earlobe, less often at the entrance of the ear canal. If the duct of this gland becomes stopped, the lubricating fatty material accumulates as a soft, rounded nodule in the skin. Infection of the cyst causes a tender abscess to form and drain. The cyst will re-form unless removed completely by surgery.

Another type of cyst occurs above the ear canal, just in front of the outer ear or, rarely, in the neck behind and below the ear. This is a remnant of the primitive gill of the early embryo, a reminder of our ancient fishy ancestors. It may appear as a tiny pitlike depression that discharges a little moisture from time to time, or a cystic swelling may develop when the opening of the pit is closed, requiring surgical removal.

Keloid of the ear. In dark-skinned people, overgrowth of scar tissue from any skin incision or injury can cause a thickened elevation on the scar called a keloid. Having the earlobes pierced for earrings sometimes results in a large, painless nodular keloid enlargement of the earlobe, harmless but unsightly. Keloids are removed surgically (see also INTEGUMENTARY SYSTEMS).

Enlargement of earlobe

Deformities. *Absence of the outer ear.* Congenital deformity or absence of the outer ear, usually on one side, sometimes on both, is often accompanied by absence of the outer-ear canal. This failure of the primitive gill structures to become properly transformed into the normal outer and middle ear is, in rare instances, hereditary. More often it occurs for no known reason. In some cases it can be traced to the damaging effects on the embryo of rubella in the mother during the first three months of her pregnancy. Since the inner ear and nerves of equilibrium and hearing come from the otic vesicle, separate from the gill structure, in most cases of deformed or absent outer ear the hearing nerve is normal. Surgical construction of a new ear canal and tympanic membrane can often improve the hearing, which has been impaired by the failure of sound conduction to reach the hearing nerve in the inner ear.

Lop ear. Lop ear, excessive protrusion of the ear from the side of the head, is a more frequent but less serious deformity of the outer ear. Surgery may be performed to bring the ears back to a more normal and less conspicuous position.

Other ailments. *Eczema.* Eczema of the skin of the outer ear, like eczema elsewhere, is an itching, scaling redness, sometimes with weeping of the affected skin. It is often the result of an allergy to a food or substance such as hair spray that comes in contact with the skin. The best treatment is discovery and avoidance of the allergen. Cortisone ointment applied topically may temporarily relieve symptoms.

Impacted earwax. The waxy substance produced by glands in the skin of the outer-ear canal normally is carried outward by slow migration to the outer layers of skin. When wax is produced too rapidly, it can accumulate, completely filling the outer-ear canal and blocking the passage of sound to the tympanic membrane, causing a painless impairment of hearing. Large plugs of earwax need to be removed by a physician. Smaller amounts may be softened by a a few drops of baby oil left in the ear overnight, then syringed out with warm water and a soft-rubber infant ear syringe.

Cancer of the outer ear. Cancer of the outer ear occurs chiefly in instances where the outer ear has been exposed for many years to direct sunlight. A small and at first painless ulcer, with a dry scab covering it, that slowly enlarges and deepens may be a skin cancer. It is diagnosed by removing a small bit of tissue from the edge and examining it under a microscope. The cancerous tissue must be completely eradicated, by either surgery or radiation, to effect a cure. Cancer that arises in the ear canal is more serious, for it may invade the bone before it is diagnosed. It is then more difficult to cure by removal. Cancers of the ear canal are rather rare, while cancers of the skin of the outer ear are more common, as well as more readily cured by removal.

MIDDLE EAR

The air-filled middle-ear cavity and the air cells in the mastoid bone that extend backward from it are supplied with air by the eustachian tube that extends from the upper part of the pharynx to the middle-ear cavity. The brain cavity lies just above and behind the middle ear and mastoid air spaces, separated from them only by thin plates of bone. The nerve that supplies the muscles of expression in the face passes through the middle-ear cavity and mastoid bone; it, too, is separated from them by only a thin layer of bone. In some instances this bony covering is incomplete, so that the facial nerve lies directly against the mucous membrane that lines the middle ear and mastoid air cells. This mucous membrane, an extension of a similar mucus-producing membrane that lines the nose and upper part of the throat, extends all the way through the eustachian tube into the middle ear and mastoid. It is subject to the same allergic reactions and infections that afflict the nasal passages. Thus, an acute head cold or other infection of the nose and throat, such as measles or scarlet fever, may extend through the eustachian tube into the middle ear and mastoid air cells. The proximity of the brain cavity to the mastoid air cells is such that an infection, if severe and untreated, may lead to meningitis (inflammation of the covering of the brain) or brain abscess. The large vein that drains blood from the brain passes through the mastoid bone on its way to the jugular vein in the neck. Infection from the middle ear can extend to this vein, resulting in "blood poisoning" (infection of the bloodstream, also called septicemia). Paralysis of the facial nerve and infection extending from the middle ear to the labyrinth of the inner ear are other possible complications of middle-ear infection. All these possibilities spring from the particular location of the small but important middle-ear cavity.

Complications of middle-ear infection

Acute middle-ear infection. Fortunately, acute middle-ear infections, called acute otitis media, are nearly always due to microorganisms that respond quickly to antibiotics. As a result, acute infection of the mastoid air cells resulting in a dangerous mastoid abscess with the possibility of meningitis, brain abscess, septicemia, infection of the labyrinth, or facial nerve paralysis, complicating an acute infection of the middle-ear cavity, has become rare. Abscess of the mastoid and the other complications of acute middle-ear infection are seen chiefly in remote regions and countries where the population lacks proper nutrition and adequate medical care.

While serious and life-threatening acute infections of

the middle ear and mastoid air cells have become rare, chronic infections, mentioned below, continue to occur, and another type of middle-ear disease, secretory otitis media, is frequent.

Secretory otitis media. In secretory otitis media the middle-ear cavity becomes filled with a clear, pale yellowish, noninfected fluid. The disorder is the result of inadequate ventilation of the middle ear through the eustachian tube. The air in the middle ear, when it is no longer replenished through this tube, is gradually absorbed by the mucous membrane, and fluid takes its place. Eventually, the middle-ear cavity is completely filled with fluid instead of air. The fluid impedes the vibratory movements of the tympanic membrane and the ossicular chain, causing a painless impairment of hearing.

The usual causes for secretory otitis media are an acute head cold with swelling of the membranes of the eustachian tube, an allergic reaction of the membranes in the eustachian tube, and an enlarged adenoid (nodule of lymphoid tissue) blocking the entrance to the eustachian tube. The condition is cured by finding and removing the cause and then removing the fluid from the middle-ear cavity, if it does not disappear by itself within a week or two. Removal of the fluid requires puncturing the tympanic membrane and forcing air through the eustachian tube to blow out the fluid. In the absence of fever and infection of the middle ear, antibiotics, which may impede the normal immune protection of the middle ear, are not necessary. In cases in which an allergic reaction is not the underlying cause of the condition, it may be necessary to insert a tiny plastic tube through the membrane to aid in reestablishing normal ventilation of the middle-ear cavity. After a time, when the middle ear and hearing have returned to normal, this plastic tube is removed. The small hole left in the tympanic membrane quickly heals.

Aero-otitis media. Aero-otitis media is a painful type of hearing loss that can result from an inability to equalize the air pressure in the middle-ear cavity when a sudden change in altitude occurs, as may happen in a rapid descent in a poorly pressurized aircraft. Allergies or a preexisting head cold may inhibit an individual's ability to equalize, which is accomplished by yawning or swallowing to open the eustachian tube. The tympanic membrane becomes sharply retracted when the air pressure becomes less within than without, while the opening of the tube into the upper part of the throat becomes pressed tightly together by the increased air pressure in the throat, so that the tube cannot be opened by swallowing. A severe sense of pressure in the ear is accompanied by pain and a decrease in hearing. Sometimes the tympanic membrane ruptures because of the difference in pressure on its two sides. More often, the pain continues until the middle ear fills with fluid or the membrane is surgically punctured. Usually aero-otitis media produced during a flight is of a temporary nature and disappears of its own accord.

Chronic middle-ear infection. Chronic infection of the middle ear occurs when there is a permanent perforation of the tympanic membrane that allows dust, water, and germs from the outer air to gain access to the middle-ear cavity. This results in a chronic drainage from the middle ear through the outer-ear canal. There are two distinct types of chronic middle-ear infection, one relatively harmless, the other caused by a dangerous bone-invading process that leads, when neglected, to serious complications.

The harmless type of chronic middle-ear disease is recognized by a stringy, odourless, mucoid discharge that comes from the surface of the mucous membrane that lines the middle ear. Medical treatment with applications of boric acid powder will dry up the chronic drainage. The perforation in the membrane may then be closed, restoring the normal structure and function of the ear with recovery of hearing.

The dangerous type of chronic middle-ear drainage is recognized by its foul-smelling discharge, often scanty in amount, coming from a bone-invading process beneath the mucous membrane. Such cases are usually caused by a condition known as cholesteatoma of the middle ear. This is an ingrowth of skin from the outer-ear canal that forms a cyst within the middle ear. An infected cholesteatoma cyst enlarges slowly but progressively, gradually eroding the bone until the cyst reaches the brain cavity, the nerve that supplies the muscles of the face, or a semicircular canal of the inner ear. The infected material within the cyst then produces a serious complication: meningitis or brain abscess, paralysis of the facial nerve, or infection of the labyrinth of the inner ear with vertigo, all of which may lead to total deafness.

Dangers of neglected cholesteatoma

Fortunately, cholesteatoma of the middle ear is now rarely so neglected as to permit development of a serious complication. By careful examination of the tympanic membrane perforation and by X-ray studies, the bone-eroding cyst can be diagnosed; it can then be removed surgically before it has caused serious harm. This operation is known as a radical mastoid or a modified radical mastoid operation. If during the same procedure the perforation in the tympanic membrane is closed and the ossicular chain repaired, the operation is known as a tympanoplasty, or plastic reconstruction of the middle-ear cavity.

Ossicular interruption. The ossicular chain of three tiny bones needed to carry sound vibrations from the tympanic membrane to the fluid that fills the inner ear may be disrupted by infection or by a jarring blow on the head. Most often the separation occurs at its weakest point, where the incus joins the stapes. If the separation is partial, there is a mild impairment of hearing; if it is complete, there is a severe hearing loss. In such a case, a hearing test demonstrates that the nerve of hearing in the inner ear is functioning normally but that sound fails to be conducted from the tympanic membrane to the inner ear. The defective ossicular chain can be surgically corrected through tympanoplasty, which allows sound to be conducted to the inner ear once again.

Otosclerosis. The commonest cause for progressive hearing loss in early and middle adult life is a disease of the hard shell of bone that surrounds the labyrinth of the inner ear. This disease of bone is known as otosclerosis, a name that is misleading, for in its early and actively expanding stage the nodule of diseased bone is softer than the ivory-hard bone that it replaces. The more appropriate name otospongiosis is sometimes used, but such is the tenacity of tradition that the older name, applied before the process was well understood, has persisted and is the term generally used.

The cause for the occurrence of the nodule of softened otosclerotic bone is unknown. There is a certain familial tendency, half the cases occurring in families in which one or several relatives have the same condition. It is one-tenth as common among blacks as among whites and twice as common in women as in men. The nodule of softened otosclerotic bone first appears in late childhood or in early adult life. Fortunately, in most cases it remains quite small and harmless, producing no symptoms, and is discoverable only if the ear bones are removed after death and examined under a microscope. Such evidence indicates that approximately 1 in 10 white adult men and 1 in 5 white adult women will be found to have such a nodule of otosclerotic bone by middle adult life.

In about 12 percent of otosclerosis cases the nodule of softened bone becomes large enough to reach the oval window containing the footplate of the stapes (stirrup). Increasing pressure caused by the expanding nodule begins to impede its vibratory movements in response to sound striking the tympanic membrane. Gradually and insidiously, affected persons begin to lose their sharpness of hearing. First they begin to lose the ability to hear faint sounds of low pitch, next they begin to have difficulty hearing the whispered voice, then they have difficulty in hearing conversation from a distance, and finally they can hear and understand the spoken voice only when it is quite loud or close to the ear. One of the characteristics of impaired hearing due to stirrup fixation by otosclerosis is retained ability to hear a telephone conversation by pressing the receiver against the head so that the sound is carried to the inner ear by bone conduction. Another characteristic of this type of impaired hearing is that hearing seems to improve while one is riding in an automobile, in a plane, or on a train. This is because the low-pitched roar of motors causes persons with normal hearing to

unconsciously raise their voices, while the individual with stirrup fixation fails to hear the low-pitched roar and thus hears better and enjoys the raised voices around him.

The diagnosis of stirrup fixation by otosclerosis is made on the basis of a history of a gradually increasing impairment of hearing with absence of any chronic infection of the middle ear or of perforation of the tympanic membrane and with hearing tests showing that the auditory nerve in the inner ear is functioning but that sound fails to be conducted properly to it. Hearing tests carried out with either a tuning fork or an audiometer demonstrate that the hearing by bone conduction is better than by air conduction.

The final and conclusive diagnosis of otosclerosis is a finding made through surgical exploration—namely, that the stapes is fixed and unable to be moved because of a nodule of bone that has grown against it. An X ray of the ear using computed tomography may be made to demonstrate that the footplate of the stapes has been invaded by otosclerosis.

Fixation of the stapes can be corrected surgically. In 1956 it was found that the fixed stapes could be removed and replaced by a plastic or wire substitute in cases in which it could not be mobilized. Today this operation, known as stapedectomy, is the one most often used to correct fixation of the stapes by otosclerosis.

Stapedectomy

The otosclerotic bone disease in some cases expands as far as the cochlea of the inner ear, causing a gradual deterioration of the auditory nerve. This progressive nerve deafness may precede, accompany, or follow fixation of the stapes. In some cases it may occur without fixation of the stapes.

INNER EAR

The labyrinth of the inner ear contains the nerve endings of the vestibular nerve—the nerve of equilibrium—and the auditory nerve, which are branches of the vestibulocochlear, or eighth cranial, nerve. The vestibular nerve ends supply the semicircular canals and the otolithic membranes in the vestibule. The auditory nerve supplies the cochlea (see above *Anatomy of the human ear: Inner ear*). Diseases of the labyrinth of the inner ear may affect both the vestibular nerve and the auditory nerve, or they may affect only the auditory nerve, with loss of hearing, or the vestibular nerve, bringing on vertigo. The commoner inner-ear diseases are described in the following paragraphs.

Nerve deafness. *Congenital nerve deafness.* Congenital nerve deafness, a defect of the auditory nerve in the cochlea, may be present at birth or acquired during or soon after birth. Usually both inner ears are affected to a similar degree, and as a rule there is a severe impairment of hearing, although in some cases of congenital nerve loss the impairment is moderate. Many cases of congenital nerve deafness have been caused by the rubella (German measles) virus in the mother during the first three months of her pregnancy, causing arrest of development of the vesicle of the embryo. This can happen during a rubella epidemic, even when the mother has no symptoms of the infection. In most cases the vestibular nerve is not affected or is affected to a lesser degree, and in most (but not all) cases the outer- and middle-ear structures are not affected. A vaccine against the rubella virus made available in 1969 has reduced the number of cases of congenital nerve deafness in developed countries.

Congenital nerve deafness acquired at or soon after birth may result from insufficient oxygen (anoxia) during a difficult and prolonged delivery or from the condition known as kernicterus, in which the baby becomes jaundiced because of incompatibility between its blood and that of the mother. In a few cases congenital nerve deafness is an inherited failure of the cochlea to develop properly. When the hearing loss is severe, speech cannot be acquired without special training. Children so afflicted must attend special classes or schools for the severely deaf, where they can be taught lipreading, speech, and sign language. Electrical hearing aids can be helpful, especially during classes, to use the remnants of hearing usually present in such cases. Another alternative, although controversial within the deaf community, is a cochlear implant, which is sometimes useful in cases of profound hearing loss or total absence of hearing when the nerve itself is present. In this operation an electrode is surgically implanted to directly stimulate the auditory nerve between the brain and the ear.

Anoxia and kernicterus

Viral nerve deafness. Viral infections can cause severe degrees of sensorineural hearing loss in one ear, and sometimes in both, at any age. The mumps virus is one of the commonest causes of severe sensorineural hearing loss in one ear. The measles and influenza viruses are less-common causes. There is no effective medical or surgical treatment to restore hearing impaired by a virus.

Effects of injury and trauma. *Ototoxic drugs.* Ototoxic (harmful to the ear) drugs can cause temporary and sometimes permanent impairment of auditory nerve function. Salicylates such as aspirin in large enough doses may cause ringing in the ears and then a temporary decrease in hearing that ceases when the person stops taking the drug.

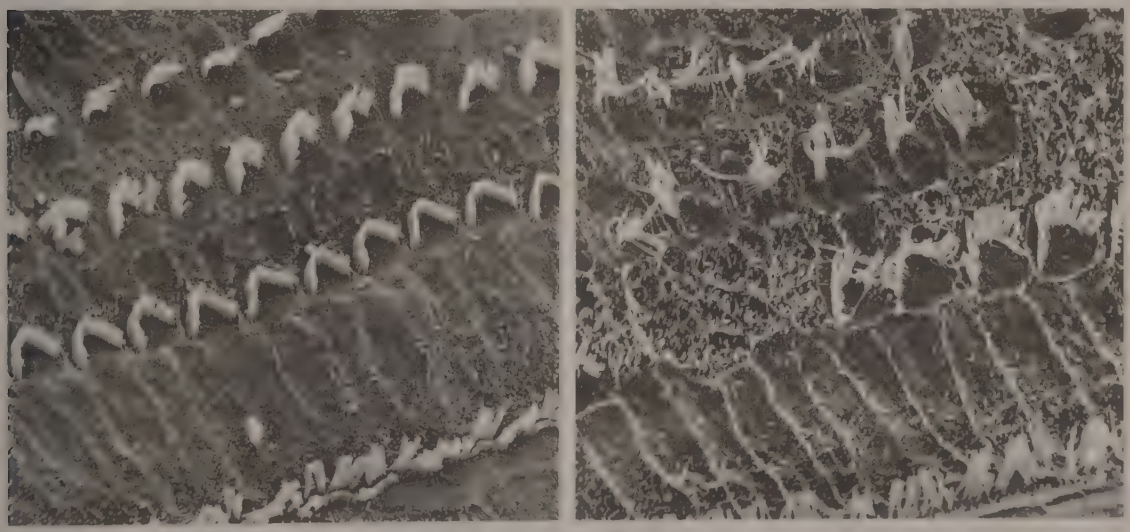


Figure 61: (Left) Portion of a healthy organ of Corti from a guinea pig shows the characteristic three rows of outer hair cells and single row of inner hair cells. (Right) Portion of a noise-damaged organ of Corti from a guinea pig exposed to sound at a 120-decibel level, similar to that experienced at a heavy metal rock concert, shows "scars" that have replaced many of the outer hair cells and shows the remaining stereocilia in disarray. Hearing is permanently damaged because lost hair cells will not be replaced, and injured cells may be dying.

Quinine can have a similar effect but with a permanent impairment of auditory nerve function in some cases. Certain antibiotics, such as streptomycin, dihydrostreptomycin, neomycin, and kanamycin, may cause permanent damage to auditory nerve function. Susceptibility to auditory nerve damage from ototoxic drugs varies greatly among individuals. In most cases, except when streptomycin is the drug taken, the more durable and less easily damaged vestibular nerve is not affected. Streptomycin affects the vestibular nerve more than it affects the auditory nerve.

Skull fracture and concussion. Skull fracture and concussion from a severe blow on the head can impair the functioning of the auditory and vestibular nerves in varying degrees. The greatest hearing loss arises when a fracture of the skull passes through the labyrinth of the inner ear, totally destroying its function.

Exposure to noise. The effects of noise exposure on hearing depend on the intensity and duration of the noise. The effects may be temporary or permanent. A single exposure to an extremely intense sound, such as an explosion, may produce a severe and permanent loss of hearing. Repeated exposures to sounds in excess of 80 to 90 decibels may cause gradual deterioration of hearing by destroying the hair cells of the inner ear, with possible subsequent degeneration of nerve fibres (Figure 61). The levels of noise produced by rock music bands frequently exceed 110 decibels. The noise generated by farm tractors, power mowers, and snowmobiles may reach 100 decibels. In the United States, legislation requires that workers exposed to sound levels greater than 90 decibels for an eight-hour day be provided some form of protection, such as earplugs or earmuffs.

Sources of high-level noise

Individuals differ in their susceptibility to hearing loss from noise exposure. Because hearing loss typically begins at the higher frequencies of 4,000 to 6,000 hertz, the effects of noise exposure may go unnoticed until the hearing loss spreads to the lower frequencies of 1,000 to 2,000 hertz. Inhalation of carbogen, a mixture of 5 percent carbon dioxide and 95 percent oxygen, for 20 minutes will accelerate recovery of hearing if administered within a few hours after excessive noise exposure.

Inflammation and tumour. *Labyrinthitis.* Labyrinthitis, an inflammation of the labyrinth of the inner ear, happens when infection occurs as a result of meningitis, syphilis, acute otitis media and mastoiditis, or chronic otitis media and cholesteatoma. Loss of both equilibrium and hearing occurs in the affected ear. Prompt antibiotic treatment sometimes arrests the damage and allows for the possibility of partial recovery of the function of the inner ear.

Acoustic neuroma. An acoustic neuroma is a benign tumour that grows on the auditory nerve near the point where it enters the labyrinth of the inner ear. The tumour causes gradual and progressive loss of auditory and vestibular nerve function on one side. Eventually the tumour grows out into the brain cavity, causing headaches and paralysis. If it is not removed, blindness and death may result. Fortunately, acoustic neuroma usually can be diagnosed early by magnetic resonance imaging (MRI) and removed before it has serious consequences.

Ménière's disease. Ménière's disease, also called endolymphatic hydrops, is a fairly common disorder of the labyrinth of the inner ear that affects both the vestibular nerve, with resultant attacks of vertigo, and the auditory nerve, with impairment of hearing. It was first described in 1861 by a French physician, Prosper Ménière. It is now known that the symptoms are caused by an excess of endolymphatic fluid in the inner ear. The diagnosis is made from the recurring attacks of vertigo, often with nausea and vomiting, impairment of hearing with a distortion of sound in the affected ear that fluctuates in degree, and a sense of fullness or pressure in the ear. The cause of the excess of endolymphatic fluid is not always known, although in many cases it results from defective functioning of the endolymphatic duct and sac, the structures that normally resorb endolymphatic fluid from the inner ear. Allergic reactions to certain foods may also cause the disease. The treatment of Ménière's disease is directed toward finding the cause of the excess of endolymphatic fluid in order to control it. If medical treatment does not relieve the repeated attacks of vertigo, surgery may be necessary.

Presbycusis. Presbycusis is the gradual decline of hearing function that results from aging. It is similar to other aging processes because it occurs at different ages and at different rates among the population. As a person ages, there is a gradual loss of cochlear hair cells, beginning at the basal end of the organ of Corti, with the result that hearing is gradually reduced and eventually lost, first for the highest audible frequencies (around 20,000 hertz) and then progressively for sounds of lower frequency. Usually the slow diminishing of hearing does not begin until after age 60. The affected individual notices increasing difficulty in hearing sounds of high pitch and in understanding conversation. Correction of a nutritional deficiency of zinc, coenzyme Q_{10}, or possibly vitamin A may stabilize the progressive hearing loss. The physician must make certain that the individual does not have a correctable impairment, such as accumulated earwax, secretory otitis media, or stirrup fixation by otosclerosis, as part of the difficulty. An electrical hearing aid is of limited help to some, while others find that a hearing aid makes voices louder but less clear and therefore is of little help. (G.E.S.)

Hearing loss in the elderly

BIBLIOGRAPHY

Animal sensory reception. *Nature and functions of sensory systems:* E.D. ADRIAN, *The Basis of Sensation: The Action of the Sense Organs* (1928), a basic work in the field of sensory physiology; *Cold Spring Harbor Symposia on Quantitative Biology,* vol. 30, *Sensory Receptors* (1965), a report of current concepts and research in the field; JOHN FIELD (ed.), *Handbook of Physiology,* section 1, *Neurophysiology,* vol. 1 (1959), a comprehensive treatise on historical as well as current concepts of sensory reception. AINSLEY IGGO (ed.), *Somatosensory System* (1973), is another comprehensive handbook. (C.A.G.W.)

Mechanoreception: JOHN FIELD (ed.), *Handbook of Physiology,* section 1, *Neurophysiology,* 3 vol. (1959–60), chapters on nonphotic receptors, posture and locomotion, vestibular mechanisms, initiation of impulses at receptors, touch and kinesis, and pain; J.D. CARTHY and G.E. NEWELL (eds.), *Invertebrate Receptors* (1968), chapters on mollusk statocysts, invertebrate proprioceptors, chordotonal organs, and mechanoreceptive transduction; P.H. CAHN (ed.), *Lateral Line Detectors* (1967), contributions of 35 investigators; M.J. COHEN and S. DIJKGRAAF, "Mechanoreception," in T.H. WATERMAN (ed.), *The Physiology of Crustacea,* vol. 2 (1961); S. DIJKGRAAF, "The Functioning and Significance of the Lateral-Line Organs," *Biol. Rev.,* 38:51–105 (1963); E. VON HOLST, "Die Arbeitsweise des Statolithenapparates bei Fischen," *Z. Vergl. Physiol.,* 32:60–120 (1950), a classical study on statoreception in the labyrinth; I.P. HOWARD and W.B. TEMPLETON, *Human Spatial Orientation* (1966), extensive coverage of the regulation of body posture; O. LOWENSTEIN, "Labyrinth and Equilibrium," pp. 60–82 in *Physiological Mechanisms in Animal Behaviour,* in *Symp. Soc. Exp. Biol.,* no. 4 (1950); D. MELLON, *The Physiology of Sense Organs* (1968); C.L. PROSSER and F.A. BROWN, *Comparative Animal Physiology,* 2nd ed. (1961), a textbook survey of mechanoreception and equilibrium; A.V.S. DE REUCK and J. KNIGHT (eds.), *Myotatic, Kinesthetic, and Vestibular Mechanisms* (1967), deals primarily with mammals; J. SCHWARTZKOPFF, "Mechanoreception," in M. ROCKSTEIN (ed.), *The Physiology of Insecta,* vol. 1 (1964). (S.Di.)

Thermoreception: J. BLIGH and H. HENSEL, "Modern Theories on Location and Function of the Thermoregulatory Centers in Mammals Including Man," in *Advances in Biometerology,* vol. 1 (1973), covers thermosensors in the central nervous system; H. HENSEL, "Physiologie der Thermoreception," *Ergebn. Physiol.,* 47:166–368 (1952), a comprehensive review with references, *Allgemeine Sinnesphysiologie: Hautsinne, Geschmack, Geruch* (1966), discusses skin receptors, with comprehensive references, and "Cutaneous Thermoreceptors," in *Handbook of Sensory Physiology,* vol. 2 (1973), describes thermoreceptors in the skin; R.W. MURRAY, "Temperature Receptors," *Advances Com. Physiol. Biochem.,* 1:117–175 (1962), covers the comparative physiology of thermoreceptors; H. PRECHT, J. CHRISTOPHERSEN, and H. HENSEL, *Temperatur und Leben* (1955), a comprehensive review on temperature and life, covering microorganisms, plants, humans, and other animals; Y. ZOTTERMAN, "Thermal Sensations," in JOHN FIELD (ed.), *Handbook of Physiology,* section 1, *Neurophysiology,* 1:431–458 (1959), and "Specific Action Potentials in the Lingual Nerve of Cat," *Skand. Arch. Physiol.,* 75:105–120 (1936), a classic showing first electrical records from specific thermosensitive nerve fibres; M. BLIX, "Experimentela bidrag till lösning av frågan om

hudnervernas specifika energi," *Uppsala LäkFör. Förh.*, 18:87–102 (1882–83), reports on the discovery of cutaneous hot and cold spots; E.A. BREARLEY and D.R. KENSHALO, "Electrophysiological Measurements of the Sensitivity of Cat's Upper Lip to Warm and Cool Stimuli," *J. Comp. Physiol. Psychol.*, 70:5–14 (1970); T.H. BULLOCK and F.P.J. DIECKE, "Properties of an Infrared Receptor," *J. Physiol.*, 134:47–87 (1956); C.B. DE WITT, "Precision of Thermoregulation and Its Relation to Environmental Factors in the Desert Iguana, Dipsosaurus Dorsalis," *Physiol. Zoöl.*, 40:49–66 (1967); H. HENSEL and K.K. BOMAN, "Afferent Impulses in Cutaneous Sensory Nerves in Human Subjects," *J. Neurophysiol.*, 23:564–578 (1960), contains first records of neural impulses from human thermoreceptors; H. HENSEL and D.R. KENSHALO, "Warm Receptors in the Nasal Region of Cats," *J. Physiol.*, 204:99–112 (1969); AINSLEY IGGO, "Cutaneous Thermoreceptors in Primates and Subprimates," *J. Physiol.*, 200:403–430 (1969); S. LANDGREN, "Convergence of Tactile, Thermal, and Gustatory Impulses on Single Cortical Cells," *Acta Physiol. Scand.*, 40:210–221 (1957); R. LOFTUS, "The Response of the Antennal Cold Receptor of Periplaneta Americana to Rapid Temperature Changes and to Steady Temperature," *Z. Vergl. Physiol.*, 59:413–455 (1968); T. NAKAYAMA *et al.*, "Thermal Stimulation of Electrical Activity of Single Units of the Preoptic Region," *Am. J. Physiol.*, 204:1122–1126 (1963), first records of impulses from thermosensors in the cat hypothalamus; D.A. POULOS and R.M. BENJAMIN, "Response of Thalamic Neurons to Thermal Stimulation of the Tongue," *J. Neurophysiol.*, 31:28–43 (1968). See also GIORGIO MACCHI, ALDO RUSTIONI, and ROBERTO SPREAFICO (eds.), *Somatosensory Integration in the Thalamus* (1983). (H.He.)

Chemoreception: J.E. AMOORE, *Molecular Basis of Odor* (1970), a technical discussion of molecular shapes and odours; M. BEROZA (ed.), *Chemicals Controlling Insect Behavior* (1970), technical reports at a symposium on pheromones and defensive secretions of insects; T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates,* 2 vol. (1965), a monumental review of invertebrate sensory physiology and neurophysiology, with extensive bibliographies; V.G. DETHIER, *The Physiology of Insect Senses* (1963), a technical review, with sections on chemoreception; H. FRINGS and M. FRINGS, *Animal Communication* (1964), a semipopular survey, including sections on chemical signaling in the animal kingdom; R. HARPER, E.C. BATE SMITH, and D.G. LAND, *Odour Description and Odour Classification* (1968), a technical review of odour theory and practical schemes of classification; T. HAYASHI (ed.), *Olfaction and Taste II* (1967), technical reports at a symposium on vertebrate chemoreception, especially electrophysiological and electron microscope studies, and discussion of theories; J.W. JOHNSTON, D.G. MOULTON, and A. TURK (eds.), "Communication by Chemical Signals," *Advances in Chemoreception,* vol. 1 (1970), a technical discussion of the field; M.R. KARE and O. MALLER (eds.), *The Chemical Senses and Nutrition* (1967), technical reports at a symposium, mostly on human chemoreception, with an extensive bibliography on taste for the years 1566–1966; W.W. KILGORE and R.L. DOUTT (eds.), *Pest Control: Biological, Physical, and Selected Chemical Methods* (1967), technical reviews by specialists, including chapters on pheromones, repellents, and antifeedants; H. KLEEREKOPER, *Olfaction in Fishes* (1969), a semipopular review especially on orientation by odours; L. MILNE and M. MILNE, *The Senses of Animals and Men* (1962), a popular survey of senses and behaviour; R.W. MONCRIEFF, *The Chemical Senses,* 3rd ed. (1967), a standard technical reference on chemoreception in vertebrates, particularly humans; G.H. PARKER, *Smell, Taste, and Allied Senses in the Vertebrates* (1922), a classic summary of earlier research and theories; H.W. SCHULTZ, E.A. DAY, and L.M. LIBBEY (eds.), *Symposium on Foods: The Chemistry and Physiology of Flavors* (1967), technical reports at a symposium, particularly on chemical analysis for odorants in foods; see *Scientific American* for excellent semipopular articles on many aspects of chemoreception (February 1964, August 1964, June 1967, May 1968, and February 1969); T.A. SEBEOK (ed.), *Animal Communication* (1968), technical reviews by specialists, with chapters on chemical signaling; E. SONDHEIMER and J.B. SIMEONE (eds.), *Chemical Ecology* (1970), technical reviews by specialists on effects of environmental chemicals on animals, including chapters on plant feeding stimulants, communication signals, defense chemicals, and fish orientation; T.H. WATERMAN (ed.), *The Physiology of Crustacea,* vol. 2 (1961), technical reviews by specialists, including chapters on senses and behaviour; V.B. WIGGLESWORTH, *The Principles of Insect Physiology,* 6th ed. (1965), a standard textbook in the field, including a chapter on chemoreception; K.M. WILBUR and C.M. YONGE (eds.), *The Physiology of Mollusca,* vol. 2 (1966), technical reviews by specialists, including chapters on chemoreception and behaviour; G.E.W. WOLSTENHOLME and J. KNIGHT (eds.), *Taste and Smell in Vertebrates* (1970), technical reports at a symposium, particularly on morphology of receptors, electrophysiology, and theories; D.L. WOOD, R.M. SILVERSTEIN, and M. NAKAJIMA (eds.), *Control of Insect Behavior by Natural Products* (1970), technical reports at a symposium particularly concerned with methods of research on feeding stimulants, deterrents, and pheromones; R.H. WRIGHT, *The Science of Smell* (1964), a semitechnical discussion of odour theories, particularly the molecular vibration theory; Y. ZOTTERMAN (ed.), *Olfaction and Taste* (1963), technical reports at a symposium, particularly on morphology, electrophysiology, and theories. Later works include H. ACKER and R.G. O'REGAN (eds.), *Physiology of the Peripheral Arterial Chemoreceptors* (1983); DIETLAND MÜLLER-SCHWARZE and ROBERT M. SILVERSTEIN (eds.), *Chemical Signals in Vertebrates: Proceedings of the Third International Symposium* (1983); D. MICHAEL STODDART (ed.), *Olfaction in Mammals: Proceedings of a Symposium of the Zoological Society of London* (1980); A.D. HASLER, A.T. SCHOLZ, and R.W. GOY, *Olfactory Imprinting and Homing in Salmon* (1983); KLAUS REUTTER, *Taste Organ in the Bullhead (Teleostei)* (1978); R.H. WRIGHT, *The Sense of Smell* (1982). (H.W.F.)

Photoreception: M.H. PIRENNE, *Vision and the Eye,* 2nd ed. (1967), optics and physiology of vision (vertebrate and invertebrate eyes) for the beginner and nonspecialist; *Handbook of Sensory Physiology:* vol. 7, pt. 1, H.J.A. DARTNALL (ed.), *The Photochemistry of Vision;* vol. 7, pt. 2, M.G.F. FOURTES (ed.), *Physiology of Photoreceptor Organs* (1971–72), authoritative treatises by leading scientists; C.G. BERNHARD (ed.), *The Functional Organization of the Compound Eye* (1966), on the optics, morphology, photochemistry, and physiology of the compound eye and other primitive eyes; GORDON L. WALLS, *The Vertebrate Eye and Its Adaptive Radiation* (1942, reprinted 1963), a classic and encyclopaedic store of knowledge of the vertebrate retina and eye; TSUNEO TOMITA, "Electrical Activity of Vertebrate Photoreceptors," *Q. Rev. Biophys.*, 3:179–222 (1970), a summary of vertebrate photoreceptor physiology; T.H. GOLDSMITH and G.D. BERNARD, "The Visual System of Insects," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta* (1972), on the visual system of insects, optics, visual pigments, and physiology; three articles on photoreceptor structure and function: M.F. MOODY, "Photoreceptor Organelles in Animals," *Biol. Rev.*, 39:43–86 (1964); and RICHARD M. EAKIN, "Lines of Evolution of Photoreceptors," in DANIEL MAZIA and ALBERT TYLER (eds.), *General Physiology of Cell Specialization* (1963), and "Structure of Invertebrate Photoreceptors," in H.J.A. DARTNALL (ed.), *Photochemistry of Vision* (1972); BRADLEY R. STRAATSMA *et al.* (eds.), *The Retina* (1969), on the morphology, function, and clinical characteristics of the vertebrate retina; HUGH DAVSON (ed.), *The Eye,* vol. 2, *The Visual Process* (1962), a textbook on photoreception and function of the visual system; three works containing research on photoreception and retinal function: *Journal of the Optical Society of America,* vol. 53, no. 1 (1963); the Cold Spring Harbor Symposia on Quantitative Biology, vol. 30, *Sensory Receptors* (1965); and the Proceedings of the International School of Physics, "Enrico Fermi," course 43, ed. by W. REICHARDT, *Processing of Optical Data by Organisms and by Machines* (1969); two reviews of visual pigment chemistry: H.J.A. DARTNALL, *The Visual Pigments* (1957); and GEORGE WALD, "Molecular Basis of Visual Excitation," *Science,* 162:230–239 (1968); T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates,* vol. 2 (1966), contains a broad survey of invertebrate sensory receptors; H.K. HARTLINE, "Visual Receptors and Retinal Interaction," *Science,* 164:270–278 (1969); FLOYD RATLIFF, *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (1965), physiological effects on vision of neural activity in the retina. Physiology of vision is also explored in JONATHAN STONE, *Parallel Processing in the Visual System* (1983); LEO M. HURVICH, *Color Vision* (1981); EBERHART ZRENNER, *Neurophysiological Aspects of Color Vision in Primates* (1983); GERALD H. JACOBS, *Comparative Color Vision* (1981). (W.H.M.)

Sound reception: There are no general texts on sound reception. Included below are some of the specialized references dealing with this subject in a technical manner.

J. SCHWARTZKOPFF, "Mechanoreception," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta,* vol. 1, pp. 509–561 (1964); M.L. WOLBARSHT, "Electrical Characteristics of Insect Mechanoreceptors," *J. Gen. Physiol.*, 44:105–122 (1960); E.G. GRAY, "The Fine Structure of the Insect Ear," *Phil. Trans. R. Soc.,* Series B, 243:75–94 (1960); C. WALCOTT and W.G. VAN DER KLOOT, "The Physiology of the Spider Vibration Receptor," *J. Exp. Zool.*, 141:191–244 (1959); E.G. WEVER and J.A. VERNON, "The Auditory Sensitivity of Orthoptera," *Proc. Natn. Acad. Sci. U.S.A.*, 45:413–419 (1959); H. FRINGS and M. FRINGS, "Uses of Sounds by Insects," *A. Rev. Ent.*, 3:87–106 (1958); R.J. PUMPHREY, "Hearing in Insects," *Biol. Rev.*, 15:107–132 (1940); "Sensory Organs: Hearing," in A.J. MARSHALL (ed.), *Biology and Comparative Physiology of Birds,* vol. 2. (1961). See also BRIAN LEWIS, *Bioacoustics: A Comparative Approach*

(1983); WILLIAM C. STEBBINS, *The Acoustic Sense of Animals* (1983); JAMES F. WILLOTT (ed.), *The Auditory Psychobiology of the Mouse* (1983). (E.G.W.)

Human sensory reception. E.G. BORING, *Sensation and Perception in the History of Experimental Psychology* (1942), a classic historical account of the early work in sensation and perception; JOHN FIELD (ed.), *Handbook of Physiology,* section 1, *Neurophysiology,* vol. 1 (1959), a technical and detailed review of modern sensory physiology; F.A. GELDARD, *The Human Senses* (1953), a scholarly overview of the senses, suitable as an introduction to the subject; D.R. KENSHALO (ed.), *The Skin Senses* (1968), a specialized report of a symposium held in 1968 that gives the reader with a special interest in this field a good idea of current research; P.M. MILNER, *Physiological Psychology* (1970), an advanced textbook in general physiological psychology with a good section on the senses; C. PFAFFMANN (ed.), *Olfaction and Taste, III* (1969), a somewhat specialized but good overview of research in olfaction and taste; J.S. WILENTZ, *The Senses of Man* (1968), an excellent popular account for the general reader that serves as a good introduction to the field. Later monographs on human sensory physiology include discussions of modern theories of pain and on space perception, in addition to traditional studies of smell, taste, vision, and hearing. See CHRISTIAAN BARNARD and JOHN ILLMAN (eds.), *The Body Machine* (1981); HERBERT HENSEL, *Thermal Sensations and Thermoreceptors in Man* (1982); TRYGG ENGEN, *The Perception of Odors* (1982); RONALD MELZACK and PATRICK D. WALL, *The Challenge of Pain,* rev. ed. (1983); LAWRENCE KRUGER and JOHN C. LIEBESKIND (eds.), *Neural Mechanisms of Pain* (1984); HERBERT L. PICK, JR., and LINDA P. ACREDOLO (eds.), *Spatial Orientation: Theory, Research, and Application: Proceedings of a Conference on Spatial Orientation and Perception, July 14–16, 1980* (1983); R. ROBIN BAKER, *Human Navigation and the Sixth Sense* (1982); MICHAEL POTEGAL (ed.), *Spatial Abilities: Development and Physiological Foundations* (1982). (C.Pf.)

Human vision: structure and function of the eye. HUGH DAVSON (ed.), *The Eye,* 4 vol. (1962; 2nd ed., vol. 1, 1969), covers the whole field of eye physiology, written by a group of experts; STEWART DUKE-ELDER *et al.* (eds.), *System of Ophthalmology,* 15 vol. in 19 (1958–76), authoritative accounts of the anatomy and physiology of the eye; E. WOLFF, *Anatomy of the Eye and Orbit,* 5th ed. (1961), the classic work on this aspect; HUGH DAVSON, *Physiology of the Eye,* 3rd ed. (1971), an account of eye physiology covering all aspects; M.H. PIRENNE, *Vision and the Eye,* 2nd ed. (1967), a simple account of certain features of eye physiology; R.A. WEALE, *The Eye and Its Function* (1960), a short and elementary account; H. VON HELMHOLTZ, *Handbuch der physiologischen Optik,* 3rd ed. (1886–96; Eng. trans., *Physiological Optics,* 3 vol., 1924–25; reprinted in 2 vol., 1962), a classic account of the psychological aspects of vision—not at all out of date, although written over 100 years ago; H.H. EMSLEY, *Visual Optics,* 5th ed., 2 vol. (1952–53), a technical account of the detailed optics of the eye. Physiological aspects of vision are discussed in HITOSHI SHICHI, *Biochemistry of Vision* (1983). Psychology of vision, including motion, depth, binocular vision, visual effects, and colour, is discussed in MARK FINEMAN, *The Inquisitive Eye* (1981).

Eye diseases and visual disorders are addressed in STEWART DUKE-ELDER (ed.), *Parsons' Diseases of the Eye,* 15th ed. (1970), a textbook for students concentrating on the more common eye conditions; F.W. NEWELL, *Ophthalmology,* 2nd ed. (1969), a standard textbook; E.S. PERKINS and P. HANSELL, *Atlas of Diseases of the Eye,* 2nd ed. (1971), illustrations of common eye conditions with brief text; D.T. VAIL, *The Truth About Your Eyes,* 2nd ed. (1959), a description for the layperson of the function of the eye and management of the more common eye diseases; F.B. WALSH and W.F. HOYT, *Clinical Neuro-Ophthalmology,* 3rd ed., 3 vol. (1969), a very detailed account of ophthalmic conditions associated with neurological diseases. A wide range of disorders and visual defects is covered in ROBERT SEKULER, DONALD KLINE, and KEY DISMUKES (eds.), *Aging and Human Visual Function* (1982); and JOHN H. DOBREE and ERIC BOULTER, *Blindness and Visual Handicap, the Facts* (1982). (E.S.P.)

Human hearing and balance: structure and function of the human ear. ANTHONY F. JAHN and JOSEPH SANTOS-SACCHI (eds.), *Physiology of the Ear* (1988), collects essays treating many different aspects of the subject; especially useful is the succinct historical account of studies of the ear and hearing by JOSEPH E. HAWKINS, JR., "Auditory Physiological History: A Surface View," pp. 1–28. A reliable and readable introductory treatise is S.S. STEVENS *et al., Sound and Hearing,* rev. ed. (1980). Although somewhat more technical, HALLOWELL DAVIS and S. RICHARD SILVERMAN, *Hearing and Deafness,* 4th ed. (1978), was also written for the nonspecialist. GEORG VON BÉKÉSY, "The Ear," *Scientific American,* 197(2):66–78 (August 1957), by a foremost research authority on the ear, describes in lay terms the mechanism of hearing, while his *Experiments in Hearing,* trans. and ed. by E.G. WEVER (1960, reprinted 1977), is the best source of information about the experimental work that won him the Nobel Prize, although it is not recommended for the novice. A well-illustrated chapter on the anatomy of the ear may be found in DON W. FAWCETT, *A Textbook of Histology,* 12th ed. (1994). For comparative anatomy from fishes to humans, the drawings in GUSTAF RETZIUS, *Das Gehörorgan der Wirbelthiere,* 2 vol. (1881–84), are still unequaled, although the work is not widely available. Other classics in the field of hearing are S.S. STEVENS and HALLOWELL DAVIS, *Hearing: Its Psychology and Physiology* (1938, reprinted 1983); E.G. WEVER, *Theory of Hearing* (1949, reissued 1970), including a good historical treatment of theories of hearing as developed through the centuries; and E.G. WEVER and MERLE LAWRENCE, *Physiological Acoustics* (1954), concerned mainly with middle-ear mechanics. A.J. HUDSPETH, "How the Ear's Works Work," *Nature,* 341(6241):397–404 (Oct. 5, 1989), gives an account of the role of hair cells in hearing. Further details are found in LEWIS G. TILNEY and MARY S. TILNEY, "Actin Filaments, Stereocilia, and Hair Cells: How Cells Count and Measure," *Annual Review of Cell Biology,* 8:257–274 (1992). As an accessible introduction to the clinical concerns of otology and audiology, JOHN BALLANTYNE, M.C. MARTIN, and ANTONY MARTIN (eds.), *Deafness,* 5th ed. (1993), remains unsurpassed. For detailed, up-to-date treatments of other topics and problems considered in this section, the series *Springer Handbook of Auditory Research* is highly recommended, especially vol. 1, *The Mammalian Auditory Pathway: Neuroanatomy,* ed. by DOUGLAS B. WEBSTER, ARTHUR N. POPPER, and RICHARD R. FAY (1992), vol. 2, *The Mammalian Auditory Pathway: Neurophysiology,* ed. by ARTHUR N. POPPER and RICHARD R. FAY (1992), vol. 7, *Clinical Aspects of Hearing,* ed. by THOMAS R. VAN DE WATER, ARTHUR N. POPPER, and RICHARD R. FAY (1996), and vol. 8, *The Cochlea,* ed. by PETER DALLOS, ARTHUR N. POPPER, and RICHARD R. FAY (1996). Information on the anatomy and physiology of the vestibular system and the disorders, peripheral and central, that can affect it may be found in the still-useful work by RALPH F. NAUNTON (ed.), *The Vestibular System* (1975). (J.E.H.)

Ear diseases and hearing disorders are discussed in MICHAEL E. GLASSCOCK III, GEORGE E. SHAMBAUGH, JR., and GLENN D. JOHNSON (eds.), *Surgery of the Ear,* 4th ed. (1990), a well-illustrated text on diseases of the ear and their surgical correction; JAMES JERGER (ed.), *Hearing Disorders in Adults* (1984); JOHN BALLANTYNE, M.C. MARTIN, and ANTONY MARTIN, *Deafness,* 5th ed. (1993); ROBERT THAYER SATALOFF and JOSEPH SATALOFF, *Hearing Loss,* 3rd ed., rev. and expanded (1993); and DAVID M. VERNICK *et al., The Hearing Loss Handbook* (1993). PHILIP H. BEALES, *Noise, Hearing, and Deafness* (1965), is a useful review in lay language of the problem of deafness and the adverse influence on hearing of excess noise exposure; a more recent text is KARL D. KRYTER, *The Handbook of Hearing and the Effects of Noise: Physiology, Psychology, and Public Health* (1994). Brief reports by panels of experts assembled by the National Institutes of Health are issued as *NIH Consensus Statements;* several statements are of clinical concern in the field of hearing and deafness: "Noise and Hearing Loss," 8(1):1–24 (Jan. 22–24, 1990), also available with the same title in *JAMA,* 263(23):3185–3190 (June 20, 1990), "Early Identification of Hearing Impairment in Infants and Young Children," 11(1):1–24 (Mar. 1–3, 1993), and "Cochlear Implants in Adults and Children," 13(2):1–30 (May 15–17, 1995), also available with the same title in *JAMA,* 274(24):1955–1961 (Dec. 27, 1995). More information on the success of cochlear implants is available in JEFFREY P. HARRIS, JOHN P. ANDERSON, and ROBERT NOVAK, "An Outcomes Study of Cochlear Implants in Deaf Patients," *Archives of Otolaryngology—Head and Neck Surgery,* 121(4):398–404 (April 1995). (G.E.S.)

Seoul

Seoul (Sŏul-t'ŭkpyŏlsi [Special City of Seoul]) was, except for a brief interregnum (1399–1405), the capital of Korea from 1394 until the formal division of the country in 1948, when it became the capital of the Republic of Korea (South Korea). Its name is derived from the ancient Korean word *sŏraböl* or *sŏböl,* meaning "capital." The city was popularly called Seoul in Korean during both the Chosŏn (Yi) dynasty (1392–1910) and the period of Japanese rule (1910–45), although the official names in those periods were Hansŏng and Kyŏngsŏng, respectively. The city was also popularly and, during most of the 14th century, officially known as Hanyang. Seoul became the official name of the city only with the founding of the Republic of Korea.

This article is divided into the following sections:

Physical and human geography 223
The landscape
The people
The economy
Administration and social conditions
Cultural life
History 225
Bibliography 225

PHYSICAL AND HUMAN GEOGRAPHY

The landscape. *The city site.* Seoul was founded in 1394 by General Yi Sŏng-gye, the founder of Korea's Chosŏn dynasty, as the capital of a unified nation. The site was a militarily defensible natural redoubt that was also an especially suitable site for a capital city, lying at the centre of an undivided Korea and adjoining the navigable Han River (Han-gang), one of the peninsula's major rivers flowing into the Yellow Sea. The contact afforded by this riverine site both with inland waterways and with coastal sea routes was particularly important to Yi because these were the routes by which grain, taxes, and goods were transported. In addition to the practical advantages, the site was well situated according to *p'ungsuchirisol,* the traditional belief in geomancy. The district chosen by Yi remains, after more than six centuries, the centre of Seoul; it is located immediately north of the Han River in the lowland of a topographic basin surrounded by low hills of about 1,000 feet (300 metres). The natural defensive advantages of the basin were reinforced two years after the city's founding by the construction of an 11-mile (18-kilometre) wall along the ridges of the surrounding hills.

Today the remains of the fortifications are a popular attraction. The old city centre is drained by a small tributary of the Han, which has been covered over by streets and expressways. Main streets and major shopping areas occupy the lower part of the basin. The original city district served to contain most of the city's growth until the early 20th century; for, although the population had grown to approximately 100,000 by the census of 1429, it had risen to only about 250,000 by the time of the Japanese annexation in 1910, almost five centuries later. The modernization program initiated by the Japanese began the first of several cycles of growth during the 20th century that extended the city limits, so that they now contain both banks of the Han River, as well as the banks of several tributary rivers. The city's boundaries now form a ragged oval about eight to 12 miles distant from the original site, except to the northwest, where they are approximately half that distance. The present boundary of Seoul is largely that established in 1963 and encompasses 234 square miles (605 square kilometres), more than twice the city area of 1948. Seoul has grown rapidly since the Korean War (1950–53). Suburbs have sprung up in the rural areas surrounding the city, and such satellite cities as Sŏngnam, Suwŏn, and Inch'ŏn have undergone considerable expansion as the capital has grown.

Growth of the city

Climate. Seoul's climate is characterized by a large annual range of temperature. The coldest month, January, has a mean temperature of 26° F (−3° C), and the warmest month, August, has a mean temperature of 78° F (25° C). Yearly precipitation in the city is about 54 inches (1,370 millimetres), with a heavy concentration during the summer months. Air pollution in the basin and in Yŏngdŭngp'o, an industrial area, has become a serious problem, caused in large part by the increasing number of automobiles and factories. For years the Han was highly polluted, but since the early 1980s pollution levels have been reduced significantly by measures to control the river's water level and by the construction of large-scale sewage treatment facilities.

The city plan. Street patterns in the city centre are basically rectangular. Streets and buildings stretch out in all directions from the old city wall's four major gates that still stand: toward Mia-dong and Suyu-dong to the north, Ch'ŏngnyang-ni to the east, Yongsan and Yŏngdŭng-p'o to the south, and Map'o and Hongje-dong to the west. Main streets, such as Ŭlchi-ro and Chong-no, are oriented east to west, but, toward the foot of the surrounding hills, topographic irregularities have some influence on the pattern. Outside the basin area of the central city, however, there are a number of radiating streets, which are interconnected by a series of circular roads. Many government office buildings are concentrated along Sejong-no, although the National Assembly building is on Yŏido island; banks, department stores, and other business offices are located along Namdaemun-no and T'aep'yŏng-no. The area of Chong-no, Myŏng-dong, and Ŭlchi-ro constitutes the central business district. The district has been transformed from an area of wooden, tile-roofed houses to one of concrete high-rise office buildings. Much of the city's expansion has been to the south of the Han, resulting in the creation of three new urban centres at Yŏido-Yŏngdŭngp'o, Yŏngdong, and Chamshil.

Housing. A shortage of housing has been a chronic problem. A number of large-scale apartment blocks were built, especially along the banks of the Han. In addition, much residential housing has been developed along the suburban fringes of the city. Old-style houses—with the traditional heated floors (*ondol*) designed for the cold winters—are still found in a few areas of the old city and adjacent to the remains of the city wall.

The people. The population of Seoul has grown extremely rapidly since 1950, and the city now has one of the highest population densities in the world. The most densely populated areas are distributed within and outside the old city and in the apartment belts along the Han. Rapid population growth in the suburbs has resulted in the creation of satellite cities around Seoul. Koreans constitute nearly all of the population, the number of foreign residents being insignificant.

The economy. *Industry and commerce.* Manufacturing, commerce, and services are the principal employers. While textile, machinery, and chemical production, food and beverage processing, and printing and publishing are still significant, the manufacture of semiconductors, computers, telecommunications equipment, and consumer electronics has grown rapidly.

Principal employers

The two most important traditional shopping areas are the extensive Tŏngdaemun (Great East Gate) Market and the smaller Namdaemun (Great South Gate) Market located near their respective gates. Comprising numerous individually owned shops, these markets serve not only Seoul but the entire country. There are also many large downtown department stores and modern shopping centres in the city.

Central Seoul and (inset) its metropolitan area.

Seoul is the centre of finance for the country. The headquarters of the major stock exchanges and banks are located there, and the city plays host to many annual trade shows.

Transportation. Although Seoul is an ancient city, it has a good road system; vast improvements have been made in the system since the Korean War, notably in widening roads and constructing of more than a dozen bridges across the Han River. Transportation facilities, however, have not been able to keep up with the demands of a large and expanding population, resulting in crowded streets and frequent traffic jams. An extensive subway system has replaced the older streetcars; this has alleviated traffic congestion somewhat and has become, with buses and railways, one of the main forms of public transport. The capital is the hub of railway lines connecting it with most provincial cities and ports, including Inch'ŏn and Pusan.

Before the Korean War, small vessels navigated up the river 37 miles to Seoul, but the demilitarized zone that now divides Korea into North and South runs partly through the mouth of the river and has deprived Seoul of its role as a river port. Hence, most goods are transported to and from the city on railways and highways. Kimp'o Airport, located in the western part of the city and long its only major airport, was joined in 2001 by Inch'ŏn (Incheon) International Airport, about 30 miles west-southwest of Seoul.

Administration and social conditions. The government consists of the Seoul Metropolitan Government, which is the executive branch, and the Seoul Metropolitan Council, the legislative body. The administrative structure contains three tiers: city, *gu* (district), and *dong* (village). The mayor of the metropolitan government and the mayors of the *gu* are elected to four-year terms. Serving under the mayors at both levels are vice mayors and directors of bureaus, offices, and divisions. The *dong* into which each *gu* is divided provide services to the residents within their administrative areas. The Seoul Metropolitan Council is headed by a president and two vice presidents, includes standing committees, special committees, and a secretariat, and has more than 100 members elected to four-year terms.

Seoul Metropolitan Council

Compulsory education applies only to the six-year elementary school, but a large proportion of elementary school graduates receive a secondary education. Most of South Korea's major universities, colleges, and research institutes are located in Seoul. (C.Le./Ed.)

Cultural life. Seoul is the country's cultural centre. It is the home of the National Academy of Arts and the National Academy of Sciences and nearly all of the nation's

Downtown Seoul seen from the Tŏksu Palace.
Woon Gu Kang

learned societies and libraries. The National Classical Music Institute, engaged in the preservation of the traditional court music of Korea and in the training of musicians, is complemented by two Western-style symphony orchestras. In addition, there are a national theatre, an opera, and a number of public and private museums, including the main branch of the National Museum of Korea on the grounds of the Kyŏngbok Palace. The Sejong Cultural Centre, to the south of the palace, has facilities for concerts, plays, and exhibitions.

Surrounded by hills, Seoul has numerous small and large parks within easy reach. Places of historical interest—including Ch'anggyŏng, Ch'angdŏk, Kyŏngbok, and Tŏksu palaces and Chong-myo Shrine—annually attract large numbers of citizens and tourists. The city also has excellent sports and recreational facilities, notably the Seoul Sports Complex, which was built for the 1986 Asian Games and the 1988 Summer Olympic Games.

HISTORY

Three Kingdoms period

The earliest historical mention of Seoul and the surrounding area dates from the 1st century BC. During the Three Kingdoms period (57 BC–AD 668) of Silla, Koguryŏ, and Paekche, the area formed a borderland between the three countries, although during the early part of the period it was most closely associated with the kingdom of Paekche. Historical accounts as well as archaeological records indicate that the original site of Paekche's capital, Wiryesŏng, was in the northeastern part of present-day Seoul. Shortly thereafter the capital was moved south across the Han River; a number of remains, including earthen walls, dwellings, and tombs, have been uncovered at that site. It was not, however, until King Munjong of Koryŏ built a summer palace in 1068 that a fairly large settlement existed on the site of the modern city.

After the formal establishment of Seoul as the capital of the unified Chosŏn (Yi) state in 1394, construction and growth were very rapid. Construction on the Kyŏngbok Palace began in 1392; it was the residence of the Yi kings from 1395 until 1592. Before residence had even been established, the construction of the city's defensive walls had been completed, although so hastily that they had to be reconstructed in 1422. The Tŏksu Palace, the construction of which began in the late 15th century, was the residence of the Yi kings from 1593 until 1611. The Ch'angdŏk Palace, begun in 1405, was the residence from 1611 to 1872, when the king moved back into the reconstructed Kyŏngbok Palace (it had been burned by the Japanese in 1592 and was not rebuilt until 1867). Throughout this period Seoul remained the centre of the "Hermit Kingdom," with little contact permitted with the outside world. The opening of Korea to diplomatic contacts with the West in 1876, at a time when the weakening Chosŏn dynasty was unable to control Western influence, led in 1905 to the establishment of a Japanese protectorate over the kingdom.

Seoul under Japanese rule

A year after the annexation of Korea to Japan in 1910, the name of the Seoul area was changed to Kyŏngsŏng, and minor changes were made in the boundary. Seoul served as the centre of Japanese rule, and modern technology was imported. Roads were paved, old gates and walls partly removed, new Western-style buildings built, and streetcars introduced.

After the end of Japanese control in 1945, Seoul came under the direct control of the central government, and in 1962 it was placed directly under the control of the prime minister. The city was left devastated by the Korean War. Out of the rubble has risen a modern city of skyscrapers and highways that has become one of the largest metropolises in the world. (C.Le.)

BIBLIOGRAPHY. The social life, customs, and ceremonies of Seoul are introduced in SEOUL METROPOLITAN GOVERNMENT, *Seoul: Traditional Ceremonies and Festivals* (1994). A historical and cultural description of the city is provided in SEOUL METROPOLITAN GOVERNMENT, *Seoul: Her History and Culture* (1992). ROBERT STOREY, *Seoul* (1999), is an informative and useful guide to the city. (Ed.)

Set Theory

Between the years 1874 and 1897, the German mathematician and logician Georg Cantor created a theory of abstract sets of entities and made it into a mathematical discipline. This theory grew out of his investigations of certain concrete problems regarding certain types of infinite sets of real numbers (see ANALYSIS: *Real analysis*). A set, wrote Cantor, is a collection of definite, distinguishable objects of perception or thought conceived as a whole. The objects are called elements or members of the set.

The theory had the revolutionary aspect of treating infinite sets as mathematical objects that are on an equal footing with those that can be constructed in a finite number of steps. Since antiquity, a majority of mathematicians had carefully avoided the introduction into their arguments of the actual infinite (*i.e.*, of sets containing an infinity of objects conceived as existing simultaneously, at least in thought). Since this attitude persisted until almost the end of the 19th century, Cantor's work was the subject of much criticism to the effect that it dealt with fictions; indeed, that it encroached on the domain of philosophers and violated the principles of religion. Once applications to analysis began to be found, however, attitudes began to change, and in the 1890s Cantor's ideas and results were gaining acceptance. By 1900, set theory was recognized as a distinct branch of mathematics.

At just that time, however, it received a severe setback through the derivation of several contradictions in its superstructure (see below *Cardinality and transfinite numbers*). The main thrust of this article is to present an account of one response to such contradictions as these. The purpose of the development here related has been to provide an axiomatic basis for the theory of sets analogous to that developed for elementary geometry. The degree of success that has been achieved in this development, as well as the present stature of set theory, has been well expressed in the Bourbaki *Éléments de mathématique:* "Nowadays it is known to be possible, logically speaking, to derive practically the whole of known mathematics from a single source, The Theory of Sets." (R.R.S./Ed.)

See the article MATHEMATICS, THE FOUNDATIONS OF for further discussion of the role and scope of set theory in the study of mathematics.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/21 and 10/22.

This article is divided into the following sections:

Introduction to set theory 226
  Fundamental set concepts 226
    Operations on sets
    Relations involved in set theory
  Essential features of Cantorian set theory 227
    Equivalent sets
    Cardinality and transfinite numbers
Axiomatic set theory 228
  Postulates of axiomatic set theory 228
    The Zermelo–Fraenkel axioms: discussion
    The Zermelo–Fraenkel axioms: formal presentation
    The Neumann–Bernays–Gödel axioms: discussion
    The Neumann–Bernays–Gödel axioms: statement
  Limitations of axiomatic set theory 231
  Present status of axiomatic set theory 231
Bibliography 232

Introduction to set theory

FUNDAMENTAL SET CONCEPTS

If the elements and sets to be considered are restricted to some fixed class of objects, such as the letters of the alphabet, the universal set (or the universe), which is commonly denoted by U, can then be defined as that which includes all of the elements—in this case, the set of all of the 26 letters. Thus, if A is one of the sets being considered, it will be understood that A is a subset of U. Another set may now be defined that includes all of the elements of U that are not elements of A. This set, which is called the complement of A, is denoted by A'. (Some writers, employing the convention of "difference sets," speak of "the complement of A with respect to U," which they denote by "$U - A$.")

Empty set

The empty (or void, or null) set, which is usually symbolized by $\emptyset$, contains no elements. One description of the empty set is that of all whole numbers that are neither even nor odd.

Operations on sets. The symbol $\cup$ is employed to denote the union of two sets. Thus, the set $A \cup B$—read "A union B" or "the union (or join) of A and B"—is defined as the set that consists of all elements belonging either to set A or set B. If sets A and B, however, have one or more members in common, their union will not duplicate those members. A committee, for example, consisting of Jones, Blanshard, Nelson, Smith, and Hixon (Committee A) may for some common purpose sit in joint session with Committee B, consisting of Blanshard, Morton, Hixon, Young, and Peters. Clearly, the union of Committees A and B must then consist of eight members rather than 10, namely, Jones, Blanshard, Nelson, Smith, Morton, Hixon, Young, and Peters.

The intersection operation is denoted by the symbol $\cap$. $A \cap B$—read "A intersect B" or "the intersection of A and B"—is defined as that set composed of all elements that belong to both A and B. Thus the intersection of the two committees in the foregoing example is the set consisting of Blanshard and Hixon.

If the set E denotes all positive even numbers and the set Q denotes all positive odd numbers, then the union of the two yields the entire sequence of positive natural numbers, and their intersection is the empty set. Any two sets the intersection of which is the empty set are said to be disjoint.

A product of two sets A and B, called a Cartesian product, is denoted by $A \times B$. This product is defined in terms of ordered pairs, analogous to the coordinates (or x and y values) of points on a Cartesian grid in analytic geometry. It is conventional to denote such pairs by enclosure in parentheses to differentiate them from unordered pairs, or sets, the members of which are enclosed in braces. In other words, the set $\{x, y\}$ is identical to the set $\{y, x\}$, but (x, y) is not the same as (y, x); two ordered pairs (a, b) and (c, d) are defined to be equal if and only if $a = c$ and $b = d$. The Cartesian product $A \times B$ may now be defined as the set consisting of all ordered pairs (x, y) for which x is an element of A and y is an element of B. An example is easily constructed: $A = \{x, y\}$, $B = \{3, 6, 9\}$, $A \times B = \{(x, 3), (x, 6), (x, 9), (y, 3), (y, 6), (y, 9)\}$.

Relations involved in set theory. The relations between sets can be of many sorts; *e.g.*, "is a subset of" ($\subseteq$), "is equivalent to" ($\sim$), "is a complement of" ($'$), "is in one-to-one correspondence with," and "has the same cardinal number as" (see below). In addition, pairing relations, defined in terms of some specific criterion, can exist between the individual elements of a set. Examples of pairing relations are: "is parallel to" ($\parallel$), "is equal to" ($=$), "is less than" ($<$), and "is the same colour as." More broadly

Pairing relations between elements

conceived, pairing can include the relations depicted on charts and graphs, on which, for example, calendar years may be paired with automobile production figures, weeks with Dow-Jones averages, and degrees of angular rotation with the lift accomplished by a cam.

The relation of one-to-one correspondence between two sets can be conceived as one in which each element of a set A is matched with an element of another set B. If $A = \{x, z, w\}$, for example, and $B = \{4, 3, 9\}$, then A is in one-to-one correspondence with B if and only if a matching such as 4 with x, 3 with z, and 9 with w obtains without any element in either set left unmatched as a remainder.

Many relations display identifiable similarities. The relations "is parallel to," "is the same colour as," and "is in one-to-one correspondence with," for example, all bear the stated relation to themselves as well as to other elements; thus, these relations are said to be reflexive. These same relations share, in addition, the property that, if an element bears the stated relation to a second element, then the second also bears that relation to the first—a property known as symmetry. Relations also have the property that, if two elements bear the stated relation to a third element, then they bear it to one another as well—a property known as transitivity.

Those relations that have all three properties—reflexivity, symmetry, and transitivity—are called equivalence relations. In an equivalence relation, all elements related to a particular element are related to each other, thus forming what is called an equivalence class. For the relation "is parallel to," for example, the equivalence class of a particular line ℓ is the set of all lines parallel to ℓ.

Equivalence classes and numbers

For each of the equivalence classes of sets, it is possible to construct an ordered set—for which not only the membership but also the sequence of its elements is significant—that can be used to name the class.

With appropriate qualifications, the cardinal of the empty set $\emptyset$ can be defined as 0; *i.e.*, $n(\emptyset) = 0$. The number 1, then, is assigned to be the cardinal of the set $\{0\}$ that contains only a single element; it is thus called the successor of 0. Similarly, the number 2, the cardinal of $\{0, 1\}$, is called the successor of 1; and 3, the cardinal of $\{0, 1, 2\}$, is the successor of 2. Continuing in this manner, the set $\mathfrak{N}$ of the natural numbers in the proper sequence $\{0, 1, 2, 3, 4, \cdots\}$ is obtained, for which the ordering is given by the successor relation.

It is this ordering that is used when one learns to count. The words one, two, three, four, $\cdots$ in proper sequence are associated with the elements in the set that are being counted. If this process stops, the set is said to be finite. Otherwise, it is said to be infinite.

There is a technical difference between cardinal and ordinal numbers. The distinction can be seen in the way that these numbers are used. A number used to designate the size of a set—*i.e.*, to answer the question, "How many?"—is used cardinally. Any use that depends on the position of the number in the prescribed sequence is the ordinal use of the number. The number found at the top or bottom of a page in a book is an example of the ordinal use of the number. (Jo.Ha./Ed.)

ESSENTIAL FEATURES OF CANTORIAN SET THEORY

At best, the foregoing description presents only an intuitive concept of a set. Essential features of the concept as Cantor understood it include: (a) that a set is a grouping into a single entity of objects of any kind; and (b) that, given an object x and a set A, exactly one of the statements "x is an element of A" (symbolized $x \in A$) and "x is not an element of A" (symbolized $(x \notin A)$) is true and the other is false. The definite relation that may or may not exist between an object and a set is called the membership relation.

Extension and abstraction

A further intent of this description is conveyed by what is called the principle of extension, viz., that a set is determined by its members: that sets A and B are equal (symbolized $A = B$) if and only if every element in A is also in B and every element in B is in A; or, in other terms, $x \in A$ implies $x \in B$ and vice versa. There exists, for example, exactly one set the members of which are 2, 5, and 7; and this set will be written by listing its elements in some order (which order is immaterial) between small braces, possibly $\{5, 2, 7\}$.

A set A is finite if for some natural number n there is a pairing of the elements of A with those of the initial segment $0, 1, \cdots, n-1$ of the natural numbers $0, 1, 2, \cdots$ in their usual order. This definition amounts to classifying a set as finite if it has a natural number n as its cardinal number (see below *Cardinality and transfinite numbers*). In principle, the brace notation is adequate for defining all finite sets.

To define infinite (*i.e.*, nonfinite) sets, Cantor used (sentential) formulas. The phrase "x is a professor" is an example of a formula; if the symbol "x" in this phrase is replaced by the name of a person, there results a declarative sentence that is true or false. The notation "$S(x)$" will be used to represent such a formula. The phrase "x is a professor at university y and x is a male" is a formula with two variables. If the occurrences of x and y are replaced by names of appropriate, specific objects, the result is a declarative sentence that is true or false. Given any formula $S(x)$ that contains the letter "x" (and possibly others) but not the letter "A," Cantor's principle of abstraction asserts the existence of a set A such that for each object x, $x \in A$ if and only if $S(x)$ holds. The unique (because of the principle of extension) set A corresponding to $S(x)$ is symbolized by $\{x \mid S(x)\}$ and read "The set of all objects x such that $S(x)$." For instance, $\{x \mid x \text{ is blue}\}$ is the set of all blue objects. This illustrates the fact that the principle implies the existence of sets the elements of which are all objects having a certain property. It is actually more comprehensive. For example, it asserts the existence of a set B corresponding to "Either x is an astronaut or x is a natural number." Astronauts have no property in common with numbers (other than both being members of B). The formula "$x \neq x$" defines the only set without elements. It is called the empty set and symbolized by $\emptyset$. The empty set is finite, for its members can be paired with those of the initial segment defined by $n = 0$.

Equivalent sets. Cantorian set theory is thus founded on the principles of extension and abstraction. To describe some results based upon these principles the notion of equivalence of sets will be defined. The set A is defined as equivalent to the set B (symbolized $A \sim B$) if and only if there exists a third set the members of which are ordered pairs such that: (a) the first member of each pair is an element of A and the second is an element of B, and (b) each member of A occurs as a first member and each member of B occurs as a second member of exactly one pair. Thus, if A and B are finite and $A \sim B$, then the third set that establishes this fact provides a pairing or matching of the elements of A with those of B. Conversely, if it is possible to match the elements of A with those of B, then $A \sim B$, because a set of pairs meeting requirements (a) and (b) can be formed (if $a \in A$ is matched with $b \in B$, then the ordered pair (a, b) is one member of the set). By thus defining equivalence of sets in terms of the notion of matching, it is formulated independently of finiteness. As an illustration involving infinite sets, $\mathfrak{N}$ may be taken to denote the set of natural numbers $0, 1, 2, \cdots$ (some authors exclude 0 from the natural numbers). Then $\{(n, n^2 \mid n \in \mathfrak{N}\}$ establishes the equivalence of $\mathfrak{N}$ and the set of the squares of the natural numbers. (R.R.S.)

Equivalence and subsets

A set B is included in, or is a subset of, a set A (symbolized $B \subseteq A$) if every element of B is an element of A. So defined, a subset may possibly include all of the elements of A, so that A can be a subset of itself. Furthermore, the empty set, because it by definition has no elements that are not included in other sets, is a subset of every set.

If every element of set B is an element of set A, but the converse is false (hence $B \neq A$), then B is said to be properly included in, or is a proper subset of, A (symbolized $B \subset A$). Thus, if A is $\{3, 1, 0, 4, 2\}$, both $\{0, 1, 2\}$ and $\{0, 1, 2, 3, 4\}$ are subsets of A; but the latter is not a proper subset. A finite set is nonequivalent to each of its proper subsets. This is not so, however, for infinite sets, as is illustrated with the set $\mathfrak{N}$ in the earlier example. (The equivalence of $\mathfrak{N}$ and its proper subset of the even natural numbers was essentially the paradox noted by Galileo in 1638.) (R.R.S./Ed.)

Infinite sets and power sets

Cardinality and transfinite numbers. The application of the notion of equivalence to infinite sets was first systematically explored by Cantor. His initial significant finding was that the set of all rational numbers (see ARITHMETIC; ANALYSIS: *Real analysis*) is equivalent to $\mathfrak{N}$ but that the set of all real numbers is not equivalent to $\mathfrak{N}$. The existence of nonequivalent infinite sets justified Cantor's introduction of transfinite cardinal numbers as measures of size for such sets. Cantor defined the cardinal of an arbitrary set A as the concept that can be abstracted from A taken together with the totality of other equivalent sets. Gottlob Frege, in 1884, and Bertrand Russell, in 1902, both mathematical logicians, defined the cardinal number $\overline{\overline{A}}$ of a set A somewhat more explicitly, as the set of all sets that are equivalent to A; this definition thus provides a place for cardinal numbers as objects of a universe whose only members are sets.

These definitions are consistent with the usage of natural numbers as cardinal numbers. A natural number, 2 for example, is first assigned to the set {0, 1} as a measure of its size; then 2 is assigned to every set equivalent to {0, 1}. Turning matters around, 2 is the concept that can be abstracted from the collection of sets equivalent to {0, 1}, or 2 may be defined as this collection of sets. Intuitively, a cardinal number, whether finite (*i.e.*, a natural number) or transfinite (*i.e.*, nonfinite), is a measure of the size of a set. Exactly how a cardinal number is defined is unimportant; what is important is that $\overline{\overline{A}} = \overline{\overline{B}}$ if and only if $A \sim B$.

To compare cardinal numbers, an ordering relation—denoted by $<$—may be introduced by means of the definition: $\overline{\overline{A}} < \overline{\overline{B}}$ if A is equivalent to a subset of B and B is equivalent to no subset of A. Clearly this relation is irreflexive $\overline{\overline{A}} \not< \overline{\overline{A}}$ and transitive $\overline{\overline{A}} < \overline{\overline{B}}$ and $\overline{\overline{B}} < \overline{\overline{C}}$ imply $\overline{\overline{A}} < \overline{\overline{C}}$.

When applied to natural numbers used as cardinals, $<$ coincides with the familiar ordering relation for $\mathfrak{N}$, so that $<$ is an extension of that relation.

The symbol $\aleph_0$ (aleph-null) is standard for the cardinal number of $\mathfrak{N}$ (sets of this cardinality are called denumerable) and $\aleph$ (aleph) is usually used for that of the set of real numbers. Then $n < \aleph_0$ for each $n \in \mathfrak{N}$ and $\aleph_0 < \aleph$.

This, however, is not the end of the matter. If the power set of a set A—symbolized $P(A)$—is defined as the set of all subsets of A, then, as Cantor proved, $\overline{\overline{A}} < \overline{\overline{P(A)}}$ for every set A—a relation that is known as Cantor's theorem. It implies an unending hierarchy of transfinite cardinals: $\overline{\overline{N}} = \aleph_0$, $\overline{\overline{P(N)}}$, $\overline{\overline{P(P(N))}}$, $\cdots$ Cantor proved that $\aleph = \overline{\overline{P(N)}}$ and was led to the question whether there is a cardinal number between $\aleph_0$ and $\aleph$, which is known as the continuum problem. A solution was completed in 1963 (see below).

There is an arithmetic for cardinal numbers based on natural definitions of addition, multiplication, and exponentiation (squaring, cubing, and so on), which deviates, however, from that of the natural numbers when transfinite cardinals are involved. For example, $\aleph_0 + \aleph_0 = \aleph_0$ (because the set of integers is equivalent to $\mathfrak{N}$), $\aleph_0 \cdot \aleph_0 = \aleph_0$ (because the set of ordered pairs of natural numbers is equivalent to $\mathfrak{N}$), and $c + \aleph_0 = c$ for every transfinite cardinal c (because every infinite set includes a subset equivalent to $\mathfrak{N}$).

The extension of the natural numbers as cardinal numbers to transfinite numbers described earlier is a typical facet of Cantorian set theory.

The Cantor paradox

The so-called Cantor paradox, discovered by Cantor himself in 1899, is the following: By the principle of abstraction, the formula "x is a set" defines a set U. It is the set of all sets. Now $P(U)$ is a set of sets and so $P(U)$ is a subset of U. By the definition of $<$ for cardinals, however, if $A \subseteq B$, then it is not the case that $\overline{\overline{B}} < \overline{\overline{A}}$. Hence, by substitution, $\overline{\overline{U}} \not< \overline{\overline{P(U)}}$. But by Cantor's theorem, $\overline{\overline{U}} < \overline{\overline{P(U)}}$. This is a contradiction. In 1902, Bertrand Russell devised another paradox of a less technical nature. The formula "x is a set and $(x \notin x)$" defines a set R of all sets not members of themselves. Using proof by contradiction, however, it is easily shown that (A) $R \in R$. But then by the definition of R it follows that (B) $(R \notin R)$. Together, (A) and (B) form a contradiction.

Axiomatic set theory

The attitude adopted in an axiomatic development of set theory is that it is not necessary to know what the "things" are that are called "sets" nor what the relation of membership means. Of sole concern are the properties assumed about sets and the membership relation. Thus, in an axiomatic theory of sets, the terms set and the membership relation $\in$ are undefined. The assumptions adopted about these notions are called the axioms of the theory. Its theorems are the axioms together with the statements that can be deduced from the axioms using the rules of inference provided by a system of logic. Criteria for the choice of axioms include: (A) their consistency (*i.e.*, that it should be impossible to derive as theorems both a statement and its negation), (B) their plausibility (*i.e.*, that they should be in accord with intuitive beliefs about sets), and (C) their richness (*i.e.*, that desirable results of Cantorian set theory can be derived as theorems).

These points are elaborated upon below.

POSTULATES OF AXIOMATIC SET THEORY

Zermelo's pioneering work

The Zermelo–Fraenkel axioms: discussion. The first axiomatization of set theory was given in 1908 by Ernst Zermelo, a German mathematician. From his analysis of the paradoxes, he concluded that they are associated with sets that are "too big," such as the set of all sets in Cantor's paradox. Thus, the axioms that Zermelo formulated are restrictive insofar as the asserting or implying of the existence of sets is concerned. As a consequence, there is no apparent way, in his system, to derive the known contradictions from them. On the other hand, the results of classical set theory short of the paradoxes can be derived. Zermelo's axiomatic theory is here discussed in a form that incorporates modifications and improvements suggested by later mathematicians, principally Thoralf Albert Skolem, a pioneer in metalogic, and Abraham Adolf Fraenkel, an Israeli mathematician. In the literature on set theory, it is called Zermelo–Fraenkel set theory (symbolized ZF), though it would seem historically more correct to call it Zermelo–Fraenkel–Skolem set theory. The 10 axioms are first discussed and then formally listed (see below *The Zermelo-Fraenkel axioms: formal presentation*).

Schemas for generating well-formed formulas. In the axioms that follow, "set" and "$\in$" are undefined terms. Lowercase Latin letters are used for variables; and variables denote sets. Equality (symbolized $=$) is taken as part of the underlying logic.

The first axiom (see axiom 1, below) conveys the idea that, as in classical set theory, a set is determined by its members. It should be noted that this is not merely a logically necessary property of equality but an assumption about the membership relation as well.

The set defined by the second axiom (see 2) is the empty (or null) set $\emptyset$.

For an understanding of the third axiom (see 3) considerable explanation is required. Zermelo's original system included the assumption (*Aussonderung* axiom) that, if a formula $S(x)$ is "definite" for all elements of a set s, then there exists a set the elements of which are precisely those elements x of s for which $S(x)$ holds. This is a version of the principle of abstraction, for it provides for the existence of sets corresponding to formulas. It restricts that principle, however, in two ways. Instead of asserting the existence of sets unconditionally, it can be applied only in conjunction with pre-existing sets. Further, only "definite" formulas (for which Zermelo offered only a vague description) may be used. Clarification was given, however, by Skolem (1922) by way of a precise definition of what will be called simply a formula of ZF. Using tools of modern logic, the definition may be made as follows:

a. For any variables x and y, $x \in y$ and $x = y$ are formulas (such formulas are called atomic).
b. If A and B denote formulas and x is any variable, then each

of the following is a formula: If A, then B; A if and only if B; A and B; A or B; not A; for all x, A; for some x, B.

Formulas are constructed recursively (in a finite number of systematic steps) beginning with the (atomic) formulas of (a) and proceeding via the constructions permitted in (b). "Not ($x \in y$)," for example, is a formula (which is abbreviated to $x \notin y$), and "There exists an x such that for every y, $y \notin x$" is a formula. A variable is free in a formula if it occurs at least once in the formula without being introduced by one of the phrases "for some x" or "for all x." Henceforth, a formula P in which x occurs as a free variable will be called a "condition on x" and symbolized $P(x)$. The formula "For every y, $x \in y$," for example, is a condition on x. It is to be understood that a formula is a formal expression—*i.e.*, a term without meaning. Indeed, a computer could be programmed to generate atomic formulas and build up from them other formulas of ever-increasing complexity using logical connectives ("not," "and," etc.) and operators ("for all" and "for some"). A formula acquires meaning only when an interpretation of the theory is spelled out; *i.e.*, when (A) a nonempty collection (called the domain of the interpretation) is specified as the range of values of the variables (thus the term set is assigned a meaning, viz., an object in the domain), (B) the membership relation is defined for these sets, (C) the logical connectives and operators are interpreted as in everyday language, and (D) the logical relation of equality is taken to be identity among the objects in the domain.

Intuitive interpretation

The terminology "a condition on x" for a formula in which x is free is merely suggestive; relative to an interpretation, such a formula does impose a condition on x. Thus, the intuitive interpretation of the third axiom schema is: given a set a and a condition on x, $P(x)$, those elements of a for which the condition holds form a set. It provides for the existence of sets by separating off certain elements of existing sets. Calling the third axiom schema an axiom schema is appropriate, for it is a schema for generating axioms—one for each choice of $P(x)$.

Axioms for compounding sets. Although the third axiom schema has a constructive quality, further means of constructing sets from existing sets must be introduced if some of the desirable features of Cantorian set theory are to be established. Each of the next three axioms is of this sort.

Using five of the axioms (see 2–6), a variety of basic concepts of classical set theory (*e.g.*, the operations of union, intersection, and Cartesian product; the notions of relation, equivalence relation, ordering relation, and function) can be defined with ZF. Further, the standard results about these concepts that were attainable in classical set theory can be proved as theorems of ZF.

Axioms for infinite and ordered sets. If I is an interpretation of an axiomatic theory of sets, the sentence that results from an axiom when a meaning has been assigned to "set" and "$\in$," as specified by I, is either true or false. If each axiom is true for I, then I is called a model of the theory. If the domain of a model is infinite, this fact does not imply that any object of the domain is an "infinite set." An infinite set in the latter sense is an object d of the domain D of I for which there is an infinity of distinct objects d' in D such that $d'Ed$ holds (E standing for the interpretation of $\in$). Though the domain of any model of the theory of which the axioms thus far discussed are axioms is clearly infinite, models in which every set is finite have been devised. For the full development of classical set theory, including the theories of real numbers and of infinite cardinal numbers, the existence of infinite sets is needed; thus the seventh axiom (see 7) is included.

The existence of a (unique) minimal set, ω, having properties expressed in the seventh axiom can be proved; its distinct members are $\emptyset$, $\{\emptyset\}$, $\{\emptyset, \{\emptyset\}\}$, $\{\emptyset, \{\emptyset\}\}$, $\{\emptyset, \{\emptyset\}\}$, $\cdots$. These elements are denoted by 0, 1, 2, 3, $\cdots$ and are called natural numbers. Justification for this terminology rests with the fact that the Peano axioms, which can serve as a base for arithmetic, can be proved as theorems. Thereby the way is paved for the construction within ZF of entities that have all the expected properties of the real numbers.

The origin of the next axiom was Cantor's recognition of the importance of being able to well-order arbitrary sets; *i.e.*, to define an ordering relation for a given set such that each nonempty subset has a least element. The virtue of a well-ordering for a set is that it offers a means of proving that a property holds for each of its elements by a process (transfinite induction) similar to mathematical induction. Zermelo (1904) gave the first proof that any set can be well-ordered. His proof employed a set-theoretic principle that he called the axiom of choice, which, shortly thereafter, was shown to be equivalent to the so-called well-ordering theorem. One form of this principle is expressed as an eighth axiom (see 8).

Well-ordering arbitrary sets

Intuitively, the axiom asserts the possibility of making a simultaneous choice of an element in every nonempty member of any set; this guarantee accounts for its name. The assumption is significant only when the set has infinitely many members. Zermelo was the first to state explicitly the axiom, although it had been used but essentially unnoticed earlier. It soon became the subject of vigorous controversy because of its unconstructive nature. Some mathematicians rejected it totally on this ground. Others accepted it but avoided its use whenever possible. Some changed their minds about it when its equivalence with the well-ordering theorem was proved as well as the assertion that any two cardinal numbers c and d are comparable (*i.e.*, that exactly one of $c < d$, $d < c$, $c = d$ holds). There are many other equivalent statements, though even today there are a few mathematicians who feel that the use of the axiom of choice is improper. To the vast majority, however, it, or an equivalent assertion, has become an indispensable and commonplace tool.

Schema for transfinite induction and ordinal arithmetic. One more axiom has been added to the list of axioms (with modifications) postulated by Zermelo. When Zermelo's eight were found to be inadequate for a fullblown development of transfinite induction and ordinal arithmetic, Fraenkel and Skolem independently proposed an additional axiom schema to eliminate the difficulty. As modified by John von Neumann, a Hungarian-born U.S. mathematician, it says, intuitively, that if with each element of a set there is associated exactly one set, then the collection of the associated sets is itself a set; *i.e.*, it offers a way to "collect" existing sets to form sets. As an illustration, each of ω, $P(\omega)$, $P(P(\omega))$, $\cdots$ is a set in the theory based on the first eight axioms. But there appears to be no way to establish the existence of the set having these sets as its members. An instance of the next schema, however, provides for its existence.

Intuitively, the ninth axiom or schema (see 9) is the assertion that if the domain of a function is a set then so is its range. That this is a powerful schema (in respect to the further inferences that it yields) is suggested by the fact that the third axiom can be derived from it and that, when applied in conjunction with the sixth axiom, the axiom of pairing can be deduced. The ninth axiom has played a significant role in developing a theory of ordinal numbers. In contrast to cardinal numbers, which serve to designate the size of a set, ordinal numbers are used to determine positions within a prescribed sequence. Following an approach conceived independently by Zermelo and von Neumann, if x is a set, the successor x' of x is the set obtained by adjoining x to the elements of x ($x' = x \cup \{x\}$). In terms of this notion the natural numbers, as defined above, are simply the succession 0, 0′, 0″, 0‴, $\cdots$; *i.e.*, the natural numbers are the sets obtained starting with $\emptyset$ and iterating the prime operation a finite number of times. The natural numbers are well-ordered by the $\in$ relation, and with this ordering they constitute the finite ordinal numbers. The axiom of infinity secures the existence of the set of natural numbers; and this set, $\aleph\omega$, is the first infinite ordinal. Greater ordinal numbers are obtained by iterating the prime operation beginning with ω. An instance of the ninth axiom or schema asserts that ω, ω', ω'', $\cdots$ form a set. The union of this set and ω is the still greater ordinal that is denoted by $\omega 2$ (employing notation from ordinal arithmetic). A repetition of this process beginning with $\omega 2$ yields the ordinals $(\omega 2)'$, $(\omega 2)''$, $\cdots$; next after all of those of this form is $\omega 3$. In this way the sequence of ordinals ω, $\omega 2$, $\omega 3$, $\cdots$

is generated. An application of the ninth axiom schema then yields the ordinal that follows all of these in the same sense in which ω follows the finite ordinals; using notation from ordinal arithmetic, it is ω^2. At this point the iteration process can be repeated. In summary, the axiom of replacement makes possible the extension of the counting process as far beyond the natural numbers as one chooses.

Cardinal and ordinal numbers

In the ZF system, cardinal numbers are defined as certain ordinals. From the well-ordering theorem (a consequence of the axiom of choice), it follows that every set a is equivalent to some ordinal number. Also, the totality of ordinals equivalent to a can be shown to form a set. Then a natural choice for the cardinal number of a is the least ordinal to which a is equivalent. This is the motivation for defining a cardinal number as an ordinal that is not equivalent to any smaller ordinal. The arithmetics of both cardinal and ordinal numbers have been fully developed. That of finite cardinals and ordinals coincides with the arithmetic of the natural numbers. For infinite cardinals, the arithmetic is uninteresting since, as a consequence of the eighth axiom, the sum and product of two such cardinals are each equal to the maximum of the two. In contrast the arithmetic of infinite ordinals is interesting and presents a wide assortment of oddities.

In addition to the guidelines already mentioned for the choice of axioms of ZF, another guideline is taken into account by some set theorists. For the purposes of foundational studies of mathematics, it is assumed that mathematics is consistent; for otherwise any foundation would fail. It may thus be reasoned that, if a precise account of the intuitive usages of sets by mathematicians is given, an adequate and correct foundation will result. Traditionally, mathematicians deal with the integers, with real numbers, and with functions. Thus an intuitive hierarchy of sets in which these entities appear should be a model of ZF. It is possible to construct such a hierarchy explicitly from the empty set by iterating the operations of forming power sets and unions in the following way.

The intuitive hierarchy of sets

The first level of the hierarchy is composed of the sequence of sets $s_0 = \emptyset$, s_1, $\cdots$, s_n, $\cdots$, in which s_{n+1} is the power set of s_n. The second level consists of the sets at the first level together with those sets obtained by iterating the power set operation any finite number of times. The third level has as its members the union of all sets constructed thus far together with those obtainable by iterating the power set operation as before. The hierarchy of sets envisaged, therefore, consists of all sets that can be obtained by proceeding to an arbitrarily large transfinite level. The domain of the intuitive model of ZF is conceived as the union of all sets in the hierarchy. In other words, a set x is in the model if it is an element of some set of the hierarchy.

Axiom for eliminating infinite descending species. From the assumptions that this system is sufficiently comprehensive for mathematics and that it is the model to be "captured" by the axioms of ZF, it may be argued that models of the first nine axioms that differ sharply from this system should be ruled out. The discovery of such a model led to the formulation by von Neumann of the tenth axiom (see 10).

This axiom eliminates from the models of the first nine axioms those in which there exist infinite descending $\in$-chains (*i.e.*, sequences $x_1, x_2, x_3, \cdots$ such that $x_2 \in x_1$, $x_3 \in x_2, \cdots$), a phenomenon that does not appear in the heuristic model described above. (The existence of models having such chains was discovered by D. Mirimanoff in 1917.) It also has other attractive consequences; *e.g.*, a simpler definition of the notion of ordinal number is possible. Yet there is no unanimity among mathematicians whether there are sufficient grounds for adopting it as an additional axiom, since it does not have the immediate plausibility that even the axiom of choice has nor has it ever been shown to have any mathematical applications.

The Zermelo–Fraenkel axioms: formal presentation.

(1) *Axiom of extension.* If a and b are sets and if, for all x, $x \in a$ if and only if $x \in b$, then $a = b$.

(2) *Axiom of the empty set.* There exists a set a such that for all x, it is false that $x \in a$.

(3) *Axiom schema of separation.* If a is a set, there exists a set b such that for all x, $x \in b$ if and only if $x \in a$ and $P(x)$. Here, $P(x)$ is any condition on x in which b is not free (it must be bound by a quantifier such as "all" or "some").

(4) *Axiom of pairing.* If a and b are sets, there exists a set (symbolized $\{a, b\}$ and called the unordered pair of a and b) having a and b as its sole members.

(5) *Axiom of union.* If c is a set, there exists a set a such that $x \in a$ if and only if $x \in b$ for some member b of c.

(6) *Axiom of power set.* If a is a set, there exists a set b such that $x \in b$ if and only if $x \in a$.

(7) *Axiom of infinity.* There exists a set a such that $\emptyset \in a$ and, if $x \in a$, then $(x \cup \{x\}) \in a$, in which $x \cup \{x\}$ is the set x with x adjoined as a further member.

(8) *Axiom of choice.* If a is a set the elements of which are nonempty sets, then there exists a function f with domain a such that for member b of a, $f(b) \in b$.

(9) *Axiom schema of replacement.* If a is a set and $B(x, y)$ a formula (in which x and y are free) such that for $x \in a$ there is exactly one y such that $B(x, y)$, then there exists a set b the members of which are the y's determined by $B(x, y)$ as x ranges over a.

(10) *Axiom of restriction.* Every nonempty set a contains an element b such that $a \cap b = \emptyset$; *i.e.*, a and b have no elements in common.

The Neumann–Bernays–Gödel axioms: discussion. The second axiomatization of set theory originated with von Neumann in the 1920s. His formulation differed considerably from ZF because the notion of function, rather than that of set, was taken as primitive. In a series of papers beginning in 1937, however, the Swiss logician Paul Bernays, a collaborator with the formalist David Hilbert, modified the von Neumann approach in a way that put it in much closer contact with ZF. In 1940, the Czech-born logician Kurt Gödel, known for his undecidability proof, further simplified the theory. This version will be called NBG.

For expository purposes it is convenient to adopt two undefined notions for NBG: class and the binary relation, $\in$, of membership (though, as is also true in ZF, $\in$ suffices). In the axioms, capital Latin letters are used as variables. For the intended interpretation, variables take classes—the totalities corresponding to certain properties—as values. A class is defined to be a set if it is a member of some class; those classes that are not sets are called proper classes. Lowercase Latin letters are used as special restricted variables for sets. For example, "for all x, $A(x)$" stands for "for all X, if X is a set, then $A(X)$"; *i.e.*, the condition holds for all sets. Intuitively, sets are intended to be those classes that are adequate for mathematics, and proper classes are thought of as those collections that are "so big" that, if they were permitted to be sets, contradictions would follow. In NBG, the classical paradoxes are avoided by proving in each case that the collection on which the paradox is based is a proper class—*i.e.*, is not a set.

Comments about the axioms that follow are limited to features that distinguish them from their counterpart in ZF. The axioms are listed later (see 11–20, below *The Neumann–Bernays–Gödel axioms: statement*).

The third axiom or schema (see 13) is presented in a form to facilitate a comparison with the third axiom schema of ZF. In a detailed development of NBG, however, there appears, instead, a list of seven axioms (not schemas) that state that for each of certain conditions there exists a corresponding class of all those sets satisfying the condition. From this finite set of axioms, each an instance of the above schema, the schema (in a generalized form) can be obtained as a theorem. When obtained in this way, the third axiom schema of NBG is called the class existence theorem.

In brief, the fourth to eighth axioms of NBG (see 14–18) are axioms of set existence. The same is true of the next axiom, which for technical reasons is usually phrased in a more general form.

Finally there may appear in a formulation of NBG an analogue (see 20) of the last axiom of ZF.

Comparison of ZF and NBG axiomatizations

A comparison of the two theories that have been formulated is in order. In contrast to the ninth schema of ZF (see 9), that of NBG (see 19) is not an axiom schema but an axiom. Thus, with the comments above about the third axiom in mind, it follows that NBG has only

a finite number of axioms. On the other hand, since the ninth axiom or schema of ZF provides an axiom for each formula, ZF has infinitely many axioms—which is unavoidable because it is known that no finite subset yields the full system of axioms. The finiteness of the axioms for NBG makes the logical study of the system simpler. The relationship between the theories may be summarized by the statement that ZF is essentially the part of NBG that refers only to sets. Indeed, it has been proved that every theorem of ZF is a theorem of NBG and that any theorem of NBG that speaks only about sets is a theorem of ZF. Finally, it has been shown that ZF is consistent if and only if NBG is consistent.

The Neumann–Bernays–Gödel axioms: statement.

(11) *Axiom of extension.* If A and B are classes and if, for all (sets) x, $x \in A$ if and only if $x \in B$, then $A = B$.

(12) Same as axiom (2).

(13) *Axiom schema for class formation.* If $P(x)$ is a condition on x in which (a) only set variables are introduced by the phrase "for all" or "for some" and (b) B is not free, then there exists a class B such that $x \in B$ if and only if $P(x)$.

(14) *Axiom of pairing.* Same as axiom (4).

(15) *Axiom of union.* Same as axiom (5).

(16) *Axiom of power set.* Same as axiom (6).

(17) *Axiom of infinity.* Same as axiom (7).

(18) *Axiom of choice.* Same as (8).

(19) *Axiom of replacement.* If (the class) X is a function and a is a set, then there exists a set b such that $y \in b$ if and only if for some x, $(x, y) \in X$ and $x \in a$; *i.e.*, the range of the restriction of a function X to a domain that is a set is also a set.

(20) *Axiom of restriction.* Every nonempty class A contains an element b such that $A \cap b = \emptyset$.

LIMITATIONS OF AXIOMATIC SET THEORY

The fact that NBG avoids the classical paradoxes and that there is no apparent way to derive any one of them in ZF does not settle the question of the consistency of either theory. One method for establishing the consistency of an axiomatic theory is to give a model; *i.e.*, an interpretation of the undefined terms in another theory such that the axioms become theorems of the other theory. If this other theory is consistent, then that under investigation must be consistent. Such consistency proofs are thus relative: the theory for which a model is given is consistent if that from which the model is taken is consistent. The method of models, however, offers no hope for proving the consistency of an axiomatic theory of sets. In the case of set theory and, indeed, of axiomatic theories generally, the alternative is a direct approach to the problem.

If T is the theory of which the (absolute) consistency is under investigation, this alternative means that the proposition "There is no sentence of T such that both it and its negation are theorems of T" must be proved. The mathematical theory (developed by the formalists) to cope with proofs about an axiomatic theory T is called proof theory, or metamathematics. It is premised upon the formulation of T as a formal axiomatic theory; *i.e.*, the theory of inference (as well as T) must be axiomatized. It is then possible to present T in a purely symbolic form; *i.e.*, as a formal language based on an alphabet the symbols of which are those for the undefined terms of T and those for the logical operators and connectives. A sentence in this language is a formula composed from the alphabet according to prescribed rules. The hope for metamathematics was that by using only intuitively convincing, weak number-theoretic arguments (called finitary methods), unimpeachable proofs of the consistency of such theories as axiomatic set theory could be given.

Gödel's theorem and the consistency of S

That hope suffered a severe blow in 1931 from a theorem proved by Gödel about any formal theory S that includes the usual vocabulary of elementary arithmetic. By coding the formulas of such a theory with natural numbers (now called Gödel numbers) and by talking about these numbers, Gödel was able to make the metamathematics of S to become part of the arithmetic of S and hence to be expressible in S. The theorem in question asserts that the formula of S that expresses (via a coding) "S is consistent" in S is unprovable in S if S is consistent. Thus, if S is consistent, then the consistency of S cannot be proved within S; rather, methods beyond those that can be expressed or reflected in S must be employed. Because, in both ZF and NBG, elementary arithmetic can be developed, Gödel's theorem applies to these two theories. Although there remains the possibility of a finitary proof of consistency that cannot be reflected in the foregoing systems of set theory, no hopeful, positive results have been obtained.

Other theorems of Gödel when applied to ZF (and there are corresponding results for NBG) assert that if the system is consistent, then (A) it contains a sentence such that neither it nor its negation is provable (such a sentence is called undecidable), (B) there is no algorithm (or iterative process) for deciding whether a sentence of ZF is a theorem, and (C) these same statements hold for any consistent theory resulting from ZF by the adjunction of further axioms. Apparently ZF can serve as a foundation for all of present-day mathematics because every mathematical theorem can be translated into and proved within ZF, or within extensions obtained by adding suitable axioms. Thus, the existence of undecidable sentences in each such theory (which entails the existence of true sentences of ZF not provable in ZF) points out the hopelessness of any attempt to base all of conceivable mathematics on a single axiomatic theory and hence implies the inadequacy of the axiomatic approach to mathematics, in particular, via axiomatic set theory.

PRESENT STATUS OF AXIOMATIC SET THEORY

The foundations of axiomatic set theory are in a state of significant change as a result of new discoveries. The situation is analogous to the 19th-century revolution in geometry, set off by the discovery of non-Euclidean geometries. It is difficult to predict the ultimate consequences of these late 20th-century findings for set theory, but already they have had profound effects on attitudes about certain axioms and have forced the realization of a continuous search for additional axioms. These discoveries have focussed attention on the concept of the independence of an axiom. If T is an axiomatic theory and S is a sentence (*i.e.*, a formula) of T that is not an axiom, and if $T + S$ denotes the theory that results from T upon the adjunction of S to T as a further axiom, then S is said to be consistent relative to T if $T + S$ is consistent and independent of T whenever both S and $\sim S$ (the negation of S) are consistent relative to T. Thus, assuming that T is consistent, if S is independent of T, then the addition of S or $\sim S$ to T yields a consistent theory. The role of the axiom of restriction (AR) can be clarified in terms of the notion of independence. If ZF′ denotes the theory obtained from ZF by deleting AR and either retaining or deleting the axiom of choice (AC), then it can be proved that if ZF′ is consistent, AR is independent of ZF′.

Relations of axiom of choice to continuum hypotheses

Of far greater significance for the foundations of set theory is the status of AC relative to the other axioms of ZF. The status in ZF of the continuum hypothesis (CH) and its extension, the generalized continuum hypothesis (GCH) are also of profound importance. (If $q(a)$ denotes "There does not exist a set b such that $\bar{\bar{a}} < \bar{\bar{b}} < \overline{\overline{P(a)}}$ the CH is $q(a)$ for $\bar{\bar{a}} = \aleph_0$ and GCH is $q(a)$ for all infinite sets a.) In the following discussion of these questions, ZF denotes Zermelo–Fraenkel set theory without AC. The first finding was obtained by Gödel in 1938. He proved that AC and GCH are consistent relative to ZF (*i.e.*, if ZF is consistent, then so is ZF + AC + GCH), by showing that a contradiction within ZF + AC + GCH can be transformed into a contradiction in ZF. In 1963, Paul J. Cohen, a U.S. mathematician, proved that (1) if ZF is consistent, then so is ZF + AC + ~CH, and (2) if ZF is consistent, then so is ZF + ~AC. Since in ZF + AC it can be demonstrated that GCH implies CH, Gödel's theorem together with Cohen's establishes the independence of AC and CH. For his proofs Cohen introduced a new method (called forcing) of constructing interpretations of ZF + AC. The method of forcing is applicable to many problems in set theory, and since 1963 it has been used to give independence proofs for a wide variety of highly technical propositions. Some of these results have opened new avenues for attacks on important foundational questions.

The current unsettled state of axiomatic set theory can be sensed by the responses that have been made to the ques-

Status of the continuum hypothesis

tion of how to regard CH in the light of its independence from ZF + AC. Someone who believes that set theory deals only with nonexistent fictions will have no concern about the question. But for most mathematicians sets actually exist; in particular, ω and $P(\omega)$ exist. Further, it should be the case that every nondenumerable subset of $P(\omega)$ either is or is not equivalent to $P(\omega)$; *i.e.*, CH either is true or is false. Followers of this faith regard the axioms of set theory as describing some well-defined reality—one in which CH must be either true or false. Thus there is the inescapable conclusion that the present axioms do not provide a complete description of that reality. A search for such axioms is in progress. One who hopes to prove CH as a theorem must look for axioms that restrict the number of sets. There seems to be little hope for this restriction, however, without changing the intuitive notion of the set. Thus the expectations favour the view that CH will be disproved. This disproof requires an axiom that guarantees the existence of more sets; *e.g.*, of sets having cardinalities greater than those that can be proved to exist in ZF + AC. So far, none of the axioms that have been proposed that are aimed in this direction (called "generalized axioms of infinity") serves to prove ~CH. Although there is little supporting evidence, the optimists hope that the status of the continuum hypothesis CH will eventually be settled. (R.R.S.)

BIBLIOGRAPHY. Overviews are provided in NICOLAS BOURBAKI, *Elements of Mathematics*, vol. 1, *Theory of Sets* (1968, reissued 1974; originally published in French, 3rd ed. rev., 1966); AZRIEL LÉVY, *Basic Set Theory* (1979), an explanation for the advanced student; and HERBERT B. ENDERTON, *Elements of Set Theory* (1972), at the advanced undergraduate or beginning graduate level. GEORG CANTOR, *Contributions to the Founding of the Theory of Transfinite Numbers*, trans. from German (1915, reissued 1955), is of historical interest. I. GRATTAN-GUINNESS (ed.), *From the Calculus to Set Theory, 1630–1910* (1980), is a historical introduction.

Various aspects of set theory are addressed by WILLIAM S. HATCHER, *Foundations of Mathematics* (1968), an overall view of axiomatic set theory and its relationship to the foundations of mathematics; PATRICK SUPPES, *Axiomatic Set Theory* (1960, reprinted 1972); GAISI TAKEUTI and WILSON M. ZARING, *Introduction to Axiomatic Set Theory*, 2nd ed. (1982); MICHAEL D. POTTER, *Sets: An Introduction* (1990), an axiomatic development of set theory; KURT GÖDEL, *The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory* (1940, reissued 1970); PAUL J. COHEN, *Set Theory and the Continuum Hypothesis* (1966), proofs of the independence of AC and CH; HERMAN RUBIN and JEAN E. RUBIN, *Equivalents of the Axiom of Choice, II* (1985); JOSEPH R. SHOENFIELD, *Mathematical Logic* (1967), a development of ZF and the independence proofs; ROBERT R. STOLL, *Set Theory and Logic* (1963, reissued 1979), an informal development of ZF; GREGORY H. MOORE, *Zermelo's Axiom of Choice: Its Origins, Development, and Influence* (1982), a good account of an important topic; ELLIOTT MENDELSON, *Introduction to Mathematical Logic*, 3rd ed. (1987), a formal development of NBG; YIANNIS N. MOSCHOVAKIS, *Descriptive Set Theory* (1980), an extensive treatment; A.S. KECHRIS and A. LOUVEAU, "Descriptive Set Theory and Harmonic Analysis," *The Journal of Symbolic Logic*, 57(2):413–441 (1992), surveying classical and modern connections with harmonic analysis; TOSHIRO TERANO, KIYOJI ASAI, and MICHIO SUGENO, *Fuzzy Systems Theory and Its Applications* (1992), a presentation of theory and of applications to various areas including image recognition and information retrieval; ANTONIO DI NOLA, *et al.*, *Fuzzy Relation Equations and Their Applications to Knowledge Engineering* (1989), an introduction; HARRY GONSHOR, *An Introduction to the Theory of Surreal Numbers* (1986), assuming some background in naive set theory; WAYNE BLIZARD, "The Development of Multiset Theory," *Modern Logic*, 1(4):319–352 (1991), and "Correction to 'The Development of Multiset Theory,'" *Modern Logic*, 2(2):219 (1991), a broad survey with emphasis on recent developments; STAN WAGON, *The Banach-Tarski Paradox* (1985, reissued 1993), a presentation of several paradoxical theorems and related problems, with a discussion of the role played by AC; M. BEKKALI, *Topics in Set Theory* (1991), developing several advanced topics; and JUDITH ROITMAN, "The Uses of Set Theory," *The Mathematical Intelligencer*, 14(1):63–69 (1992), demonstrating applications to areas of mathematics not generally considered to be closely related to set theory. (Jo.Ha./R.R.S./Ed.)

Sex and Sexuality

Sex is represented through the sum of features by which members of species can be divided into two groups—male and female—that complement each other reproductively. For most forms of life, sexual reproduction is essential for both propagation and long-term survival. Exceptions exist: not all sexual behaviour necessarily leads to reproduction, and reproduction in some species occurs nonsexually—that is, without sexual contact.

The article is divided into the following sections:

Animals and plants 233
- Sexual and nonsexual reproduction 233
  - Sex cells
  - Sexuality: complementary mating types
- The adaptive significance of sex 234
  - Reproduction and evolution
  - Life cycles adjusted to environmental change
- The origin of sex and sexuality 235
  - Differentiation of the sexes
  - Mating
  - Courtship
- Sex patterns 238
  - Sex differences in animals
  - Seasonal or periodic sexual cycles
- Sex determination 239
  - Sex chromosomes
  - Abnormal chromosome effects
  - Parthenogenesis
  - Effects of environment
  - Hormones

Human beings 240
- Types of behaviour 241
  - Solitary behaviour
  - Sociosexual behaviour
- Physiological aspects 242
  - Sexual response
  - Genetic and hormonal factors
  - Nervous system factors
  - Development and change in the reproductive system
- Psychological aspects 244
  - Effects of early conditioning
  - Sexual problems
- Social and cultural aspects 246
  - Social control of sexual behaviour
  - Class distinctions
  - Economic influences
  - Legal regulation
- Sexually transmitted diseases 249
  - Common sexually transmitted organisms
  - Acquired immunodeficiency syndrome
- Homosexuality 249
  - Classification and prevalence
  - Prenatal and environmental influences
  - Homosexual lifestyles

Bibliography 252

Animals and plants

SEXUAL AND NONSEXUAL REPRODUCTION

Because the life span of all individual forms of life, from microbes to man, is limited, the first concern of any particular population is to produce successors. This is reproduction, pure and simple. Among lower animals and plants it may be accomplished without involving eggs and sperm. Ferns, for example, shed millions of microscopic, nonsexual spores, which are capable of growing into new plants if they settle in a suitable environment. Many higher plants also reproduce by nonsexual means. Bulbs bud off new bulbs from the side. Certain jellyfish, sea anemones, marine worms, and other lowly creatures bud off parts of the body during one season or another, each thereby giving rise to populations of new, though identical, individuals. At the microscopic level, single-celled organisms reproduce continually by growing and dividing successively to give rise to enormous populations of mostly identical descendants. All such reproduction depends on the capacity of cells to grow and divide, which is a basic property of life. In the case of most animals, however, particularly the higher forms, reproduction by nonsexual means is apparently incompatible with the structural complexity and activity of the individual.

Nonsexual reproduction

Although nonsexual reproduction is exploited by some creatures to produce very large populations under certain circumstances, it is of limited value in terms of providing the variability necessary for adaptive advantages. Such so-called vegetative forms of reproduction, whether of animals or plants, result in individuals that are genetically identical with the parent. If some adverse environmental change should occur, all would be equally affected and none might survive. At the best, therefore, nonsexual reproduction can be a valuable and perhaps an essential means of propagation, but it does not exclude the need for sexual reproduction.

Sexual reproduction not only takes care of the need for replacement of individuals within a population but gives rise to populations better suited to survive under changing circumstances. In effect it is a kind of double assurance that the race or species will persist for an indefinite time. The great difference between the two types of reproduction is that individual organisms resulting from nonsexual reproduction have but a single parent and are essentially alike, whereas those resulting from sexual reproduction have two parents and are never exact replicas of either. Sexual reproduction thus introduces a variability, in addition to its propagative function. Both types of reproduction represent the capacity of individual cells to develop into whole organisms, given suitable circumstances. Sex is therefore something that has been combined with this primary function and is responsible for the capacity of a race to adapt to new environmental conditions.

Sex cells. The term sex is variously employed. In the broad sense it includes everything from the sex cells to sexual behaviour. Primary sex, which is generally all that distinguishes one kind of individual from another in the case of many lower animals, denotes the capacity of the reproductive gland, or gonad, to produce either sperm cells or eggs or both. If only sperm cells are produced, the reproductive gland is a testis, and the primary sex of the tissue and the individual possessing it is male. If only eggs are produced, the reproductive gland is an ovary, and the primary sex is female. If the gland produces both sperm and eggs, either simultaneously or successively, the condition is known as hermaphroditic. An individual, therefore, is male or female or hermaphrodite primarily according to the nature of the gonad.

Gonads and primary sex

As a rule, male and female complement each other at all levels of organization: as sex cells; as individuals with either testes or ovaries; and as individuals with anatomical, physiological, and behavioral differences associated with the complemental roles they play during the whole reproductive process. The role of the male individual is to deliver sperm cells in enormous numbers in the right place and at the right time to fertilize eggs of female individuals of the same species. The role of the female individual is to deliver or otherwise offer eggs capable of being fertilized under precise circumstances. In the case of hermaphrodite organisms, animal or plant, various devices are employed to ensure cross-fertilization, or cross-pollination, so that full advantage of double parentage is obtained. The basic requirement of sexual reproduction is that reproductive

cells of different parentage come together and fuse in pairs. Such cells will be genetically different to a significant degree, and it is this feature that is essential to the long-term well-being of the race. The other sexual distinctions, between the two types of sex cell and between two individuals of different sex, are secondary differences connected with ways and means of attaining the end.

Sexuality: complementary mating types. The complementarity of both male and female sex cells and male and female individuals is a form of division of labour. Male sex cells are usually motile cells capable of swimming through liquid, either freshwater, seawater, or body fluids, and they contribute the male cell nucleus but little else to the fertilization process. The female cell also contributes its nucleus, together with a large mass of cell substance necessary for later growth and development following fertilization. The female cell, however, is without any capacity for independent movement.

In other words, small male cells (sperm cells, spermatozoa, or male gametes) are burdened with the task of reaching a female cell (egg, ovum, or female gamete), which is relatively large and awaits fertilization. A full complement of genes is contributed by both nuclei, representing contributions by both parents, but, apart from the nucleus, only the egg is equipped or prepared to undergo development to form a new organism. A comparable division of labour is seen in the distinction between male and female individuals. The male possesses testes and whatever accessory structures may be necessary for spawning or delivery of the sperm, and the female possesses ovaries and what may be needed to facilitate shedding the eggs or to nurture developing young. Accordingly there is the basic sex, which depends on the kind of sex gland present, and sexuality, which depends on the different structures, functions, and activities associated with the sex glands.

The role of each sex

THE ADAPTIVE SIGNIFICANCE OF SEX

When two reproductive cells from somewhat unlike parents come together and fuse, the resulting product of development is never exactly the same as either parent. On the other hand, when new individuals, plant or animal, develop from cuttings, buds, or body fragments, they are exactly like their respective parents, as much alike as identical twins. Any major change in environmental circumstances might exterminate a race since all could be equally affected. When eggs and sperm unite, they initiate development and also establish genetic diversity among the population. This diversity is truly the spice of life and one of the secrets of its success; sex is necessary to its accomplishment.

In each union of egg and sperm, a complete set of chromosomes is contributed by each cell to the nucleus of the fertilized egg. Consequently, every cell in the body inherits the double set of chromosomes and genes derived from the two parental cells. Every time a cell divides, each daughter cell receives exact copies of the original two sets of chromosomes. The process is known as mitosis. Accordingly, any fragment of tissue has the same genetic constitution as the body as a whole and therefore inevitably gives rise to an identical individual if it becomes separated and is able to grow and develop. Only in the case of the tissue that produces the sex cells do cells divide differently, and genetic differences occur as a result.

During the ripening of the sex cells, both male and female, cell divisions (known as meiosis) occur that result in each sperm and egg cell having only a single set of chromosomes. In each case the set of chromosomes is complete—*i.e.*, one chromosome of each kind—but each such set is, in effect, drawn haphazardly from the two sets present in the original cells. In other words, the single set of chromosomes present in the nucleus of any particular sperm or egg, while complete in number and kinds, is a mixture, some chromosomes having come from the set originally contributed by the male parent and some from the female. Each reproductive cell, of either sex, therefore contains a set of chromosomes different in genetic detail from that of every other reproductive cell. When these in turn combine to form fertilized eggs or fertile seeds, the double set of chromosomes characteristic of tissue cells is reestablished, but the genetic constitution of all such cells in the new individual will be the same as that of the fertilized egg—two complete sets of genes, randomly derived from sets contributed by the two different parents. Variation is thus established in two steps. The first is during the ripening of the sex cells, when each sperm or egg receives a single set of chromosomes of mixed ancestry. None of these cells will have exactly the same combination of genes characteristic of the respective parent. The second step occurs at fertilization, when the pair of already genetically unique sex cells fuse together and their nuclei combine, thus compounding the primary variation.

Reproduction and evolution. Sexual reproduction appears to be a process serving two opposing needs. The individuals produced must be almost exactly like their parents if they are to succeed; *i.e.*, to grow and reproduce in turn, under the prevailing circumstances. At the same time they should exhibit a wide range of differences so that some at least can survive under different environmental circumstances. The first business of reproduction is to produce perfect working copies of the parental organism, without any mistakes. The second is to introduce novelties—*i.e.*, new models that make possible other life styles. Extreme conservatism, in either sexual or nonsexual reproduction, may be disastrous to the species in the long run. Extreme variability may also be detrimental, resulting in the production of too high a percentage of misfits. A delicate balance has to be struck. Variability is necessary but must be kept within bounds. Sex is responsible for controlled diversity, without which adaptation and evolution could not take place.

Significance of reproduction

Natural selection operates in two ways on this basic diversity inherent in any particular population or community. In a stable environment, where there is little change during a long period of time, except for the regular diurnal and seasonal changes, those individuals most likely to survive and produce offspring are those that are most like their parents at all stages of their existence. The more radical departures from the established types fail either to grow or to compete successfully and consequently do not reproduce. The less radical departures struggle along but leave progeny in proportionately smaller numbers. If, however, a significant long-term change occurs in the environment, the established types are likely to suffer, while other types that previously had been weeded out now may be favoured. They may become the more successful at surviving and growing and consequently replace themselves more readily than do others. They, in turn, become the establishment, and the older type is jeopardized. A constant interplay persists between a changeable environment and a variable population. This is adaptation. If environmental change continues in the same general direction, adaptation also continues in the initial direction, and eventually significant evolution becomes apparent.

The variability or diversity resulting from sexual reproduction is vital in two ways. It permits the process of natural selection to work and allows a population of organisms to adapt to new conditions. It also serves as a corrective mechanism. During nonsexual reproduction, particularly of single-cell organisms, large populations of virtually identical individuals are readily built up and maintained for a great many generations. Sooner or later, however, more and more abnormalities appear and, usually, a general waning of vigour ensues. When such organisms subsequently fuse together in pairs, equivalent to sexual reproduction, a rejuvenation and reestablishment of healthy strains generally follows.

Life cycles adjusted to environmental change. Both sexual and nonsexual reproduction may be exploited or adjusted to meet widely fluctuating environmental conditions, especially those of a regular seasonal character. This phenomenon is particularly striking in the case of the smaller or simpler forms of animal and plant life that have a life-span of a year or less. The seeds of annual plants germinate in the spring, grow and set seed in turn during the summer, and die in the fall. Only the sexually produced seeds persist and represent the species during the long winter season. Certain small, though common, freshwater creatures have a similar cycle. The microscopic eggs of

Alteration of sexual and nonsexual phases

Hydra and of *Daphnia*, for example, lie at the bottom of ponds throughout the winter, each within a tough protective case. In late winter or early spring, a new generation of hydras develops, each individual becoming attached to a stone or vegetation and feeding on small crustaceans by means of its long slender tentacles. The daphnias, or so-called water fleas, emerge at about the same time and grow rapidly to maturity. In both cases the growing season, usually from spring until fall, is a time for intensive reproduction by whatever means is most effective. Hydras bud off new hydras continually, each new hydra repeating the process, with the size of populations limited only by available food. Only late in the season, when the food supply drops off and the temperature drops, does the riotous splurge of nonsexual reproduction come to an end. Then each individual ceases to bud and produces either minute ovaries or testes, and in some species, both. Eggs become fertilized, encased, drop into the mud, and await the coming of the following spring, while the parental creatures die as living conditions worsen with approaching winter. Such is a general pattern of life, widely seen among creatures whose individual existence is measured in weeks or months but whose race must persist in some form at all times if extinction is to be avoided.

So it is with *Daphnia* and many other organisms. The *Daphnia* also changes according to the times, but it alternates between one form of sexual reproduction and another. Sexually, the *Daphnia* is exquisitely adapted to the little world in which it lives. Under ideal conditions every member of a *Daphnia* community is female. All those first hatching out from winter eggs in the spring are females. Each produces a succession of broods during the month or two of its individual existence, all offspring being females. Each such female, generation after generation, during the spring and summer seasons, produces eggs that develop at once without need or opportunity of being fertilized. No males in fact are present. Every individual is a self-sufficient breeding female. Population explosions occur wherever environmental circumstances are favourable. Eventually, however, conditions inevitably change for the worse, either because of effects inherent in any population explosion or because every season comes to an end. Food becomes scarce because of too many consumers; space becomes crowded and in some degree polluted; chilly days succeed the warmth of summer. Whatever the cause, and well before disaster can strike, the creatures respond in remarkable ways. On the first signal that conditions may be getting less than good, a certain number of the eggs produced by a population of *Daphnia* develop into males, each with testes in place of ovary, together with certain secondary sexual characteristics. A scattering of males through the virgin paradise, however, is only the first step, a preparedness in case conditions go from bad to worse. If there has been a false alarm, the females continue to produce female-producing eggs that develop parthenogenetically—that is, without benefit of fertilization—and the males die off without performing any sexual function. But if the environmental signal means the beginning of the end of congenial conditions, a cell in the ovary of each female grows to form a larger egg than usual, and it is of a type that must be fertilized. Then mating between the sexes takes place, and the resulting special, fertilized eggs become thickly encased and alone survive the winter season after becoming separated from the parent.

Parthenogenesis

Wherever small aquatic creatures live in bodies of water that may freeze in winter or dry up in summer, similar adaptations may be seen in many forms of life besides hydras and water fleas. Certain small fish, known as the annual fishes, have individual life-spans of about six months. The life-span itself is in fact adapted to the period during which active existence is possible in their particular habitat. When the water holes, swamps, and puddles in which they live begin to dry up, mating takes place, and the fertilized eggs drop into the mud. The parents die, and the eggs remain in a state of suspended development until the next rainy season occurs. The race must continue whatever the circumstances, and all sex is directed toward this end.

THE ORIGIN OF SEX AND SEXUALITY

All sexual reproduction, no matter how large or small the organisms may be, is a performance of single cells. Only at the level of single cells can the essential genetic recombinations be accomplished. So in every generation new life begins with the egg, which is a single cell, however large it may be. Egg and sperm unite at fertilization, but the fertilized egg is as much a single cell as before. When did it all begin? The generally accepted answer is that the fundamental, or molecular, basis of sexuality is an ancient evolutionary development that goes back almost to the beginning of life on earth, several billion years ago, for it is evident among the vast world of single-celled organisms, including bacteria.

In these lowest forms of life, sex and reproduction are distinct happenings. Reproduction is accomplished in most cases entirely by fission, which is simply cell division repeated regularly, as long as the environmental conditions permit. As long as crowding and other adverse changes are avoided, cells divide, and the daughter cells grow and divide again, for weeks or months on end. This process occurs in both plantlike and animal-like single-celled organisms and in bacteria as well. Under certain other conditions, such cell organisms come together and fuse in pairs, a form of sexual behaviour at its primary level and comparable to the fusion of an egg and sperm. In all such case, a combined cell is produced in which nuclear exchange or recombination has occurred. Pairing off of this sort takes place sooner or later in all forms of unicellular life, even where no outwardly distinguishable differences can be detected between the pairing individuals. The lack of discernible differences between the members of mating pairs, however, does not mean that pairing occurs between identical individuals. In the much investigated *Paramecium* and other protozoan organisms, two separate populations of cells may continue to increase almost indefinitely by ordinary cell division of single individuals, but when two such populations are mixed together, mating generally occurs immediately between individuals from the two different sources. The fusion, or pairing, has essentially the same function as the fusion of the male and female nucleus during the process of fertilization of eggs of higher forms. It is the basis of sex, the essential event in all cases being the genetic or chromosomal recombination.

Individual mating cells (*i.e.*, eggs, sperm, or even whole single-celled organisms) may be called gametes whether or not they are distinguishable from one another. Yet even among the varius single-celled organisms, mating commonly occurs between individuals of two different kinds. This kind of mating is seen most often among the single-celled organisms known as flagellates. In some species the gametes may be alike and all are motile, progressing through the water by means of one or more whiplike flagella similar to the tail of a sperm. In other species, all individuals may still be motile, but pairing occurs between individuals of different sizes. In still others, one of the two mating types may be very small and motile, and the other, large, with stored nutritional material, and nonmotile. All degrees of differentition between male and female gametes can be found, and it is probable that the basic and characteristic distinction between the sex cells of both animal and plant life in general was established very early in the course of evolution, during the immense period of time when virtually all living organisms consisted of single cells.

Kinds of gametes

This division of labour between mating types, male and female, respectively, is nature's way of attaining two ends. These are the bringing together of the gametes so that fusion may take place and the accumulation of reserves so that development of a new organism can be accomplished. The first calls for as many motile cells as possible; the second calls for cells as large as possible. These different requirements are practically impossible to satisfy by a single type of cell. Accordingly, and especially in multicelled animals of all sorts, male gametes, or spermatozoa, are extremely small, extremely motile, and are produced in enormous numbers. The larger the number, the greater the likelihood that some will encounter and fertilize eggs. On the other hand, the female gametes, or ova, individually need to be as large as possible since the larger the size

and the more condensed the internal nutritional reserves, the farther along the path of embryonic development the egg can travel before hatching must occur and the new organism must fend for itself. Nevertheless, eggs in general are caught between the desirability of being individually as large as they can be and the persisting need to be produced in reasonably large numbers, so that an assortment of differing indivuduals is produced from a single pair of parents. A large number of offspring ensures that a proportion, at least, will survive the environmental hazards faced by all developing organisms in some degree.

Differentiation of the sexes. Animals and plants, apart from microscopic kinds of life, consist of enormous numbers of cells coordinated in various ways to form a single organism, and each consists of many different kinds of cells specialized for performing different functions. Certain tissues are set aside for the production of sexual reproductive cells, male or female as the case may be. Whether they are testes or ovaries or, as in some animals and plants, both together in the same parental individual, they are typically contained within the body, and therefore the sex cells usually need to be passed to the outside in order to function. Only in certain lowly creatures such as hydras is there a simpler state, for in hydras the testes and ovaries form in the outermost layer of cells of the slender, tubular body, and the sex cells when ripe burst directly from the simple, bulging gonads into the surrounding water. With few other exceptions, in all other creatures the gonads are part of the internal tissues and some means of exit is necessary. In some, such as most worms, all that is needed are small openings, or precisely placed pores, in the body wall through which sperm or eggs can escape. In most others, more is needed and a tubular sperm duct or an oviduct leads from each testis or ovary, through which the sex cells pass to the exterior. This is minimal equipment, except where none is needed. The gonad and its duct is accordingly comparable to other glands in the body; that is, the gland is generally a more or less compact mass of cells of a particular, specialized kind, together with a duct for passage of the product of the tissue to the site of action. Gonads secrete—*i.e.*, produce and transmit—sex cells that usually act outside the body.

Differentiation between the sexes exists, therefore, as the primary difference represented by the distinction between eggs and sperm, by differences represented by nature of the reproductive glands and their associated structures, and lastly by differences, if any, between individuals possessing the male and female reproductive tissues, respectively.

Sex cells, sexual organs, other sexual structures, and sexual distinction between individuals constitute a series of evolutionary advances connected with various changes and persisting needs in the general evolution of animals and, to some degree, of plants as well. In other words, no matter how large or complex a creature may become, it still needs to deliver functional sex cells to the exterior. This condition is almost always the case for sperm cells. Among aquatic animals, particularly marine animals whose external medium, the ocean, is remarkably similar chemically to the internal body fluid medium of all animals, eggs are also in most cases shed to the exterior, where development of the fertilized eggs can proceed readily. Even so, time and place are important. Starfish, sea urchins, and many others, for instance, accumulate mature eggs and sperm in the oviducts and sperm ducts until an appropriate time when all can be shed at once. When one member of a group of such creatures begins to spawn, chemicals included in the discharge stimulate other members to do the same, so that a mass spawning takes place. One might say that the more they are together the more variable their offspring may be. This situation actually is the crux of the matter for nearly all forms of life, because while it may be possible for a single individual to possess both male and female gonads, producing both sperm and eggs, it remains generally desirable, if not essential, that eggs be fertilized by sperm produced by another individual. Cross-fertilization results in a much greater degree of variability than does self-fertilization. The existence of two types of individuals, male and female, is the common means of ensuring that cross-fertilization will be accomplished, since then nothing else is possible. Where the sexes are separate, therefore, all that is necessary is that members of the opposite sex get together at a time and place appropriate for the initial development of fertilized eggs. Typically, spawning of this sort is a communal affair, with many individuals of each sex discharging sex cells into the surrounding water. This process is only suitable, however, when eggs are without tough protective cases or membranes; that is, only when eggs are readily fertilizable for some time after being shed and while drifting in the sea. In this circumstance there is no need for individuals of the opposite sex to mate in pairs, nor is such mating practiced.

The value of cross-fertilization

Mating. Mating between two individuals of the opposite sex becomes necessary when eggs must be fertilized at or before the time the eggs are shed. Whenever eggs have a protective envelope of any kind through which sperm cannot penetrate, fertilization must take place before the envelope is formed. The envelope may at first be a gluey liquid, which covers the egg and solidifies as a tough egg case, as in all crustaceans, insects, and related creatures. It may be a thick membrane of protein deposited around the egg, as in fishes generally; or it may be a material that swells up as a mass of jelly surrounding the eggs after the eggs have been shed, as in frogs and salamanders. And finally, it may be a calcified shell, as in birds and reptiles. In all of these organisms the sperm must reach the egg before the protective substance is added, except in those forms in which a small opening or pore persists in the egg membrane through which sperm can enter.

When and how such eggs need to be fertilized depends on the nature of the protective membranes and the time and place of their formation. The jelly surrounding frog and toad eggs, for instance, swells up immediately after the eggs are shed. Mating and fertilization must take place at the time of spawning. Male frogs mount the back of female frogs and each clasps his mate firmly around the body, which not only helps press the egg mass downward but brings the cloacal opening of male and female close together. Eggs and sperm are shed simultaneously, and the eggs are fertilized as they leave the female body. Fish eggs are also fertilized as or shortly after they are shed, although fish have no arms and mating generally is usually no more than a coming together of the two sexes side by side, so that simultaneous shedding of sperm and eggs can be accomplished. In other creatures the mating procedure may be much more complicated, depending on various circumstances. Crustaceans such as crabs and lobsters, for example, mate in somewhat the same manner as frogs, with the male holding on to the female by means of clawlike appendages and depositing sperm at the openings of the oviducts, which are typically situated near the middle of the undersurface of the body.

Mating modifications imposed by the land environment. Greater problems arise on land than in water. Eggs produced by truly terrestrial creatures are either retained in the parental body during their development or must be fully protected from drying up. Protective membranes must be tough indeed. More importantly, however, sperm cells must still be deposited where they can swim toward the eggs, for they cannot survive or function except in a watery solution of dilute salts. In all terrestrial creatures, except those that return to water to breed, sperm can survive only in the body of the male or female organism. All insects, therefore, must mate in order for eggs to be fertilized, and all have appendages at the rear of the body that serve as a copulatory device capable of being used even when in flight. Sperm is injected into the female's duct or storage sac, either for immediate fertilization or for later use. The queens of bees, ants, and termites, in fact, mate once and for all during a nuptial flight and thereafter use the stored sperm to fertilize all the eggs they subsequently produce.

The land vertebrates have to cope with much the same breeding circumstances as the insects. Man is more aware of these procedures because they happen mostly in much larger creatures and also because he has some fellow feeling for them. Reptiles, birds, and even the most primitive surviving mammals—namely, the platypus and spiny anteater of Australasia—produce yolky eggs encased in a

Necessity of internal fertilization in land animals

more or less rigid calcareous shell. Moreover, within the shell, a thick layer of albumen surrounds the egg proper. Both the albumen and the shell are added after the ovum leaves the ovary and during its passage down the oviduct. Fertilization must take place, if at all, as the eggs enter the oviduct, for neither the albumen nor the shell can be penetrated by spermatozoa. Sperm must therefore be introduced into the female and must be able to make their way up to the end of the oviduct, which is a very long journey for so small a cell. An enormous number must begin the journey to make sure that some will reach the goal.

Sexual anatomy. In reptiles and birds of both sexes, as in amphibians and fish, a single opening to the exterior serves jointly for both the intestine and reproductive duct. This is the cloaca, or vestibule. Nevertheless, copulation of a sort occurs in all three groups of terrestrial vertebrates: the reptiles, birds, and mammals. With the exception of man, the male always mounts the female from the rear or back, and in both reptiles and birds the cloacal openings are pressed closely together to form a continuous passage from one individual to the other. With one exception, the archaic tuatara (*Sphenodon*) of New Zealand, all present-day reptiles have an erectile penis, derived from the cloacal wall, that delivers the sperm into the proper duct. One mating may serve for a long time, and there are cases known in which female snakes have laid fertile eggs after months and sometimes years of isolation in captivity. On the other hand, a penis of any sort is lacking in most kinds of birds, and the pressing together of the cloacal apertures seems to serve well enough. The most advanced copulatory procedure is that of mammals. In mammals the cloaca has become replaced by separate openings for the reproductive duct and intestine, respectively. Eggs have become microscopic, devoid of shell, yolk, and virtually all albumen, although they still need to be fertilized as they enter the upper end of the oviduct. A well-developed, erectile penis is always present in the male for the ejaculation of stored sperm well up the reproductive passage of the female. Accordingly, the two sexes have become strikingly differentiated anatomically, with regard to delivery of sperm, compared with the seemingly primitive anatomical equipment of birds.

Courtship. The coming together of two members of the opposite sex is a necessary preliminary to mating. It may be accomplished by two individuals independently of any larger congregation, or it may result from two individuals pairing off within a breeding population that may have assembled even from the ends of the earth. In the one the problem is to find one another; in the other the problem is to find the appropriate place, called the staging area. In both cases timing and some sort of navigation are important. Mass assembly appears to be the more effective, although a local crowd of any kind of animal may be an open invitation to predators, human or otherwise, and may on occasion become disastrous.

The searching out of a solitary individual by another of the opposite sex can be a difficult matter. In the dark depths of the ocean, for instance, where fish and other marine life forms are extremely scarce and scattered, the chance of encounter is rare indeed. The small angler fish (*Photocorynus spiniceps*) that cruise around at great depths are most unlikely to meet a member of the opposite sex at a time or place when the female happens to be ready to shed her eggs. As a form of insurance to this end, however, any small, young male that happens to meet a large female, apparently at any time, immediately fastens on to her head or sides by his jaws and thereafter lives a totally parasitic existence sustained by the juices of the female body. Sperm thus becomes available at any time the female may produce eggs to be fertilized.

On land this individual procedure of searching out is common among insects and the more predatory mammals. Male crickets and cicadas sound their familiar signals, by night or by day, which attract any females within hearing distance. More remarkable are those insects and other creatures that produce living light, in some cases for no apparent purpose but in others, such as the firefly, for signalling between the sexes in the dark of summer nights. The male individuals, always more dispensable than females, fly freely at considerable risk, flashing their light at regular intervals. The light of the female, perched more safely on some tall grass, winks back as though it were a landing light, and so they come together. Each of the several species of firefly has its own flash code, or rhythm, and any wasteful attempt at interspecies mixing is avoided. On the same principle, female moths send their personal perfume into the night air, and those males that detect the scent fly toward the source, the winner taking all. Mammals also depend mainly on their sense of smell, being generally colour blind, not too attentive to sound, and, apart from the grazing and browsing creatures, mainly active at night. The scented sex appeal of a cat in heat, whether domestic or wild, excites all the males in the neighbourhood and, with or without the sound of voice, male and female come quickly together in the dark. In all of these, courting is mostly uncalled for since only ready-to-mate individuals are involved in this sexual searching in the dark.

Sexual attractants

Courting is necessary whenever the male is a supplicant. A female may not be ready to mate, and stimulation in the form of dance or song may be required to create the mood; or, as is commonly the case, there is a surplus of available and eager males, and one must be chosen among many. However it may be, courting is most practiced not only when the female is in command of the final outcome but also when the mating procedure presents certain difficulties. A small male spider dances before a larger and ever ravenous female in an effort to induce her sexual interest rather than her hunger. Birds especially, however, depend on courtship as a preliminary to mating. The mating of birds represents copulation in its simplest form, without benefit of significant anatomical devices. Bird wings are a poor substitute for arms in a sexual embrace. Consequently the fullest cooperation between male and female is essential to success. In most birds a long-lasting, often lifetime, bonding becomes established between a male and female, a bonding that is usually reinforced by ritual behaviour at certain intervals, particularly during the onset of each breeding season and on various occasions when the individuals meet after short periods of separation. In some species a new mate may be taken each season or, as in sparrows, a general promiscuity may prevail.

One important aspect of courtship concerns the question of recognition. In gull colonies, for instance, members of the opposite sex look very much alike, and, at least to humans, the various individuals of one sex or the other may appear exactly the same. The advantages, with regard to successful production, incubation, and rearing of eggs and young, of permanent or semipermanent mate selection, however, are as great in gull colonies as elsewhere. The preliminaries to such a mutual selection not only establish a bond, by various posturings, but also establish the many small idiosyncrasies of action that add up to individuality and make one bird distinguishable among many within a colony, at least to its mate.

Many different forms of sex-oriented behaviour have consequently evolved among birds, depending on the character and particular needs of the various species. Penguins apparently not only look alike to human observers but also to themselves. Penguins seemingly have trouble even distinguishing between the sexes. Being unable to dance or sing, though they can make a lot of noise, male penguins can do little more than offer a pebble to a prospective female. If she accepts it as a token contribution to nest making, the match is on. If it is rejected, the suitor may have picked an unready female or even another male. In the case of most birds, however, the male can either sing, particularly the smaller kinds, or can strut and dance, with wings and feathers displayed, and some species, such as the lyre bird, continue to enchant the female by sight and sound together. In general, the need for physical mating has led to courtship and an emotional bonding between mating pairs throughout much of the animal kingdom at the higher level, particularly among birds and mammals. These are primarily utilitarian functions relating to the survival of the species, but in their fullest expression they represent what seem to man to be among the finest attributes of life.

SEX PATTERNS

Since the great value of sex as distinct from reproduction is the reassortment and recombination of genes every generation, sex cells from two separate parents ordinarily give rise to the greatest variaton, unless the parental individuals are themselves too closely related to each other. The presence of male and female individuals, respectively, generally produced in approximately equal numbers, is characteristic of so much of the animal kingdom that it appears to be the natural state. All that is certain, however, is that this condition has evolved as the most effective means to the particular end, and it may have done so independently among the various more or less unrelated groups of animals. The condition of separate sexes is not a universal fact, and two sexes within the same individual is typical of the more sluggish or actually attached kinds of animal life. Earthworms, slugs, land snails, flatworms, tapeworms, barnacles, sea squirts, and some others are all double-sexed individuals, or hermaphrodites. All have ovaries and testes producing mature eggs and sperm at the same time. Nevertheless, cross-fertilization is accomplished, and self-fertilization, even though possible, is generally avoided. Of those kinds of animal life mentioned above, all except the sea squirts have well-encased eggs that need to be fertilized before being laid. Mutual copulation, whereby each member of a mating pair of individuals introduces sperm into the body of the other member, is characteristic of these creatures, with the exception of the sea squirts.

Hermaphrodites

When animals shed sperm and comparatively naked eggs into the surrounding water, as is the case in sea squirts, self-fertilization is difficult to avoid. Most creatures have evolved an effective separation of the sexes between different individuals. Even so, there are more ways than one of accomplishing this. The common means is to produce male and female individuals that are constitutionally different, yet an equally effective procedure is for all individuals to be constitutionally the same but to become mature as male or female at different stages of the growth cycle. The oyster on its rock changes sex from male to female and back again once or twice a year. Certain shrimps also are hermaprodites. Each young shrimp of this kind grows up to be a male and is fully and functionally a male when about half the size of the females. As the next season approaches, his testes shrink, no more spermatozoa are produced, and ovaries begin to enlarge. As full growth is reached, the shrimp that had been a male becomes a typical female, ready to mate again, but this time with a young male of a newer generation. The system works as well as any other and clearly has its points. In fact the hagfish, not a true fish but a more primitive jawless vertebrate, also changes sex regularly, from year to year.

Sex differences in animals. In many animals, sexual differences are apparent in addition to the primary sex differentiation into males with testes and females with ovaries and apart from the accessory structures and tissues associated with the presence of one kind of sex gland or the other. Secondary sex differentiation in sexually distinct individuals is to be seen in many forms. In humans, for example, the beard and deep voice of the male and the enlarged breasts of the female are features of this sort. The great claw of the fiddler crab, the antlers of a moose, the great bulk and strength of a harem master in a fur seal colony, the beautiful fan tail of the peacock, and the bright feathers of other birds, are all distinctively male characteristics, and all are associated with the sexual drive of males. Females, by and large, are of comparatively quiet disposition and relatively drab appearance. Their function is to produce and nurture eggs, as safely and usually as inconspicuously as possible. The male function is to find and fertilize the female, for which both drive and display are generally required.

It is the business of sperm to be active and so find an egg. Similarly it is the busines of males to find a female and mate with her if possible. The male drive, or male eagerness, is a consequence of this special function of males. In nature, males possessing a strong eagerness to mate will find more females and leave more progeny than males lacking in sex drive. The progeny moreover will tend to inherit the drive of the parent. Males therefore are generally competitive with other males, with a premium placed on physical strength and sex drive and also on various devices for the attraction and stimulation of the female. The various exclusively male features already listed are all examples of characteristics of this sort, and they are related to the securing of female mates rather than the actual fertilization of eggs or to the problems of survival and adaptation.

Seasonal or periodic sexual cycles. In most animals sexual reproduction is seasonal or rhythmical, and so is sexual behaviour, whether in the form of courtship, drive, or other activities that lead to mating. In the marine fireworm of the West Indies, for instance, individuals of both sexes live in crevices on the sea floor but come out to breed where their fertilized eggs can drift and develop in the water above. But they can only find one another by means of the luminescence they themselves produce, which is an eerie light visible only in complete darkness. Each spring or summer month they emerge and swim to the surface about one-half hour after sunset when all daylight is gone but only before the moon can rise, a situation that confines them to a monthly breeding period of three or four days after the full of the moon. They follow a lunar rhythm. So do the grunion, a common fish along the southern California coast. Here again mating takes place when all is dark and the tide is high. Pairing occurs in the wash of the waves on the sand; fertilized eggs become immediately buried and there develop until the next high spring tides reach and wash the upper level sand nearly two weeks later. The mysterious biological clocks that apparently all living things possess adjust the rhythms of life to the needs of the particular organism. Some of these timing processes call internal signals on a regular day and night basis; others, on a somewhat longer cycle that keeps pace with the moon rather than the sun; and many, especially in the larger animals, run on a seasonal, or annual, cycle. Many activities are brought into line with the regular changes occurring in the environment. Sex and reproduction, however, are adjusted mainly with regard to two functions; namely, safety while mating, which is therefore commonly in the dark, and the launching of the new generation at a time or season when circumstances are most favourable.

Biological control of sexual cycles

Birds lay eggs, and most mammals deliver their young in early spring, when the months ahead are warm and food is plentiful. Sex for the most part is adjusted to this end. Among the mammals, for example, the period of development within the womb varies greatly, from less than three weeks in the smallest to almost a year in the largest and certain others. Yet with few exceptions, the time for birth is in the spring. The time for mating in most cases is accordingly adjusted to this event: the larger the offspring at birth, the earlier the mating must take place. The horse and the great whales mate in spring and deliver in spring; roe deer mate in summer and deliver in spring; goat and sheep mate in the fall and deliver in spring. Even the elephant, which has a 22-month pregnancy, delivers in spring but must mate in early summer two years before. In small creatures, however, such as mice, rats, hamsters, and shrews, where the gestation, or pregnancy, period is about three weeks, reproduction is still seasonal, but there is time during the warmer months for several broods to be conceived and raised. In others, expediency may prevail, and mating may occur at a time to suit the convenience of the pairing animals. The little brown bat, for instance, mates in the fall, and yet ovulation does not take place until winter has passed; the spermatozoa survive the winter in the uterus and fertilize the eggs when they in turn arrive there five or six months later. In some other creatures mating occurs at a convenient time, eggs are fertilized, but development itself is suspended at an early stage for a time so that hatching or birthing, depending on the kind of animal, takes place when circumstances are suitable.

In all of this, the time of the mating season is clearly regulated, both with regard to the physiological condition of the animal and to the environmental conditions. The urge and capacity to mate depends on the ripeness of the gonads, male or female. In most animals, the reproductive glands wax and wane according to the seasons; that is, with

Hormonal regulation of cycles

an annual rhythm or else with a shorter cycle. Hormones are mainly in control of this rhythm. Sex hormones, male or female, respectively, are produced by the gonads themselves and cause or maintain their growth and at the same time cause the various secondary sexual characteristics of the male or female individual to become enhanced. Male hormone increases masculinity, even when injected into a female. Female canaries injected with male hormone no longer behave as females and shortly begin to sing loud and long and commence the courtship activities of a male. A hen thus injected grows a larger comb, starts to crow, and begins to strut.

The production of these hormones is in turn controlled by hormones of the pituitary gland. Pituitary hormones stimulate ovarian or testicular tissue, which secretes the sex hormones. The sex hormones not only maintain the growth of the sexual tissues generally but inhibit the secretion of pituitary hormones, so that the process does not get out of hand. The pituitary activity, however, is also influenced by external conditions, particularly by stimuli received indirectly from light. The annual growth of ovaries or testes that occurs in late winter and early spring in frogs, reptiles, birds, and mammals is initiated by the steadily increasing period of daylight. In response to this changing day length, female frogs are packed with eggs and male frogs are ready to croak by the time the mating period arrives. The large eggs of reptiles and birds are ready to be fertilized, and the males are showing whatever they may have to display at the proper time. In mammals, the female comes into heat, the uterus undergoes the preparatory changes for taking care of fertilized eggs, and the male usually has but one thought in his mind. But as daylight ceases to lengthen, the sexual drive slowly diminishes.

SEX DETERMINATION

The determination of the sex of an individual, with regard to both the primary sex—*i.e.*, whether the ovaries or the testes develop—and the various secondary sexual characteristics may be rigorously controlled from the start of development or may be subject to later influences of a hormonal or environmental nature. However this may be, in order to appreciate the action of the control systems, the point of departure is that animals were primitively hermaphrodite, that during early stages of evolution every individual probably possessed both male and female gonads. Differentiation into separate sexes, each possessing male or female gonads but not both at the same time, is a device to ensure cross-fertilization of eggs, whether this is accomplished by having the two types of sexual gland mature at different stages of the growth of the individual, as in some shrimp and others, or whether by the production of two distinct types of individuals, as in most species of animals. This point of view is important because the question ceases to be how testes are caused to develop in the male organism and ovaries in the female but how, in a potentially double-sexed organism, the development of one or the other sex is suppressed. That such is the case is seen as clearly as anywhere in the human condition itself. Neither sex is completely male or female. Females have functional, well-developed mammary glands. Males also have mammary glands, undeveloped and nonfunctional although equipped with nipples. Males have a penis for delivering sperm, but females have a small, nonfunctional equivalent—the clitoris. These are secondary sexual features, to be sure, but the difference between the sexes is in the degree of their development, not a matter of absolute presence or absence.

Early development of reproductive systems in young animals

The basis for this is seen in the very beginnings of the development of the reproductive system, in frog, mouse, and man alike. In the young embryo a pair of gonads develop that are indifferent or neutral, showing no indication whether they are destined to develop into testes or ovaries. There are also two different duct systems, one of which can develop into the female system of oviducts and related apparatus and the other into the male sperm duct system. As development of the embryo proceeds, either the male or the female reproductive tissue differentiates in the originally neutral gonad of the mammal.

In the frog and other lower vertebrate animals, the picture is even clearer. The original gonad consists of an outer layer of cells and an inner core of cells. If the individual is to be a male, the central tissue grows at the expense of the outer layer. If it is to be a female, the outer tissue grows at the expense of the central core tissue. If both should grow, which is a possibility although a rare occurrence, the individual will be a hermaphrodite. Anything that influences the direction taken therefore may be said to determine sex.

Sex chromosomes. In most species of animals the sex of individuals is determined decisively at the time of fertilization of the egg, by means of chromosomal distribution. This process is the most clear-cut form of sex determination. When any cell in the body divides, except during the formation of the sex cells, each daughter cell receives the full complement of chromosomes; *i.e.*, copies of the two sets of chromosomes derived from the sperm cell and egg, respectively. The two sets are similar except for one pair of chromosomes. These are the so-called sex chromosomes, and the pair may be exactly alike or they may be obviously different, depending on the sex of the individual. The sex chromosomes are of two types, which are designated X and Y, and the pair of sex chromosomes may consist of two X chromosomes or of an X and Y paired together. In mammals (including man) and flies, the cells of males contain an XY pair and the cells of females contain an XX pair. On the other hand, in butterflies, fishes, and birds, the cells of females contain an XY pair and those of males contain an XX pair. In either case the Y chromosome is generally smaller than the X chromosome and may even be absent. What is most important concerning chromosomal sex determination is whether the cells of the individual contain one X chromosome or two X chromosomes. Human beings, for example, have cells with 22 pairs of nonsexual chromosomes, or autosomes, together with an XX pair or an XY pair. The female has a total of 46 functional chromosomes; the male has 45 plus a Y, which is mainly inert. Sex determination thus becomes a matter of balance. With one X chromosome plus the 44 autosomes in every cell, the whole course of development of primary and secondary sexual characteristics is toward the male; with two X chromosomes plus the autosomes in every cell, the whole system is swung over to the female.

The manipulation of this control system is readily accomplished during the special process of cell division that takes place in the gonads to produce sperm and eggs and their subsequent union at fertilization. In mammals, for example, since all cells in the female contain two X chromosomes, all the eggs will receive a single X chromosome when they are formed. All eggs are accordingly the same in this respect. In contrast, all cells in the male have the XY constitution, and therefore, when the double set of chromosomes is reduced to a single set during the formation of the spermatozoa, half of the spermatozoa will receive an X and half will receive a Y. Consequently, when an egg is fertilized by a sperm, the chances are about equal that the sperm will carry an X or will carry a Y, since the two types are inevitably produced in equal numbers. If it carries an X, the XX female constitution results; if a Y, then the XY male constitution results.

Abnormal chromosome effects. Occasionally, however, the processes of chromosomal reassortment and recombination occurring during sex cell formation and fertilization depart somewhat from the normal course. Sperm and eggs may be produced that are oversupplied or undersupplied with sex chromosomes. Fertilized eggs in humans may, for instance, have abnormal sex chromosome constitutions such as XXX, XXY, or XO. Those with the triple-X chromosome constitution have all the appearance of normal females and are called, in fact, superfemales, although only some will be fertile. Those with the XO (one X, but lacking Y altogether) constitution, a much more common condition, are also feminine in body form and type of reproduction system but remain immature. Individuals with the XXY constitution are outwardly males but have small testes and produce no spermatozoa. Those with the more abnormal and relatively rarer constitutions XXXXY and XXYY are typically mentally defective and in the latter

case are hard to manage. Thus abnormal combinations generally result in an infertility on the one hand and an abnormal sexuality in the whole system, for either too little or too much of what is ordinarily good can be disastrous.

Very different kinds of abnormal development resulting from faulty chromosomal distribution are particularly observable in insects. The most common form in flies is an individual that is male on one side, female on the other, with a sharp line of demarcation. In other cases one-quarter of the body may be male and three-quarters female, or the head may be female and the rest of the body, male. These types are known as gynandromorphs, or sexual mosaics, and result from aberration in the distribution of the X chromosomes among the first cells to be formed during the early development of the embryo. This condition is unknown among higher animals.

Sexual mosaics

Parthenogenesis. The unfertilized, ripe egg possesses all the potentiality for full development. The process of fertilization by a spermatozoon introduces the nucleus of the male sex cell into the female egg, a process that increases the differences between parent and offspring and may determine the sex of the new individual and also stimulates the egg to begin development. These two functions are separate. Parthenogenetic development, without benefit of sperm, occurs naturally in various kinds of animals besides the waterflea (*Daphnia*), already described. Artificial, or experimental, parthenogenesis is readily brought about in many other species and by a variety of means. Mature, unfertilized eggs of starfish, sea urchins, various worms, and other marine invertebrate animals can be caused to develop by treatment with a weak organic acid. Unfertilized frog eggs can be readily caused to develop by gentle pricking of the egg surface with the tip of a fine glass needle that has been dipped in lymph. In nature the eggs of various creatures can develop with or without the aid of spermatozoa. The sex of parthenogenetically developed individuals, insofar as it depends on the chromosomal constitution of the developing egg, is consequently affected. Frog eggs developing parthenogenetically become males, since only one X chromosome is present in each cell. In nature, where varying conditions call for various responses, the system is usually more complicated, although based on the general relationship that individuals with the XX constitution will be female and those with a single X will be males. A queen honeybee, for instance, begins her reproductive life with a store of sperm received from a male during her nuptial flight. Throughout spring and summer almost all eggs become fertilized and develop into females (either as nonfertile female workers or as new fertile queens, depending on the nature of food received during growth). Toward the end of summer, when the sperm supply runs low, eggs cease to be fertilized and, when laid, develop into drones, ready to mate with a new queen should occasion arise. In other cases, even parthenogenetically developing eggs may become female individuals through a process of chromosome doubling, which takes place in the mature but unfertilized eggs. Thus certain wasps, waterfleas, and others are able to produce many exclusively female generations in succession.

Effects of environment. Sex chromosomes, however, do not determine sex directly but do so through their control of such cell activities as metabolism and hormone production. Their determinative influence, indirect though it is, may be complete. On the other hand, environmental conditions may play the dominating role. In the case of *Bonellia*, a unique kind of marine worm, all eggs develop into small larvae of a sexually indifferent kind. Those that settle freely on the sea floor grow into comparatively large females, each of which has a long, broad extension, the proboscis, at its front end. Those larvae that happen to settle on the proboscis of a female, however, fail to grow beyond a certain minute size and become dwarf males, permanently attached to the female body. The sex-determining factor appears to be the environmental carbon dioxide tension, which is relatively high at the surface of living tissue.

Hormones. Because in most developing animals the reproductive gland is essentially neutral to begin with, there is generally some possibility that agents external to the gland, particularly chemical agents—*i.e.*, hormones—circulating in the blood system, may override the sex-determining influence of the sex chromosomes. In the chick, for example, the sex can be controlled experimentally by such means until about four hours after hatching. If a female chick is injected on hatching with the male sex hormone, testosterone, it will develop into a fully functional cock. Even when injected at later stages of growth, the male hormone causes extra early growth of the comb, crowing, and aggressive behaviour after being injected in either male or female chicks. Female sex hormones, such as estrogen, on the other hand, stimulate early growth of the oviduct in the female and feminize the plumage and suppress comb growth when injected in the male.

This susceptibility of the reproductive glands, and sexuality in general, to the influence of sex hormones is particularly acute in mammals, where the egg and embryo, unprotected by any shell, develop in the uterus exposed to various chemicals filtering through from the maternal blood stream. A developing embryo eventually produces its own sex hormones, but they are not manufactured in any quantity until the anatomical sex of the embryo is already well established. One of the curious things about sex hormones, however, is that the reproductive glands are not the only tissues that produce them. The placenta, through which all exchange between fetus and mother takes place, itself produces tremendous amounts of female sex hormone, together with some male hormone, which are excreted by the mother during pregnancy. This condition is true of humans, as well as of mice and rats. As a rule these hormones are produced too late to do any harm, but not always. The female embryo is fairly immune inasmuch as additional female hormone merely causes a child to be more feminine than usual at an early age. Male embryos, however, may be seriously affected if the female hormone catches them at an early stage. Boy babies may be born that are truly males but under the impact of the feminizing hormone appear superficially to be females and are often raised as such. As a rule, even when older, they have more or less sterile, undescended testes; an imperfect penis; well-developed breasts; an unbroken voice; and no beard. One in a thousand may be like this and on occasion may have won in women's Olympic competitions. In other cases, those somewhat less severely affected, during adolescence when the hidden testes begin to secrete their own male hormones in abundance, the falsely female characteristics become suppressed, and the voice, beard, breasts, and sexual interest take on the pattern of the male. What were thought to be girls in their youth change into the men they were meant to be upon reaching maturity. (N.J.B.)

Production and effects of sex hormones in mammals

Human beings

Human sexual behaviour may be defined as any activity—solitary, between two persons, or in a group—that induces sexual arousal. There are two major determinants of human sexual behaviour: the inherited sexual response patterns that have evolved as a means of ensuring reproduction and that are a part of each individual's genetic inheritance, and the degree of restraint or other types of influence exerted on the individual by society in the expression of his sexuality. The objective here is to describe and explain both sets of factors and their interaction.

It should be noted that taboos in Western culture and the immaturity of the social sciences for a long time impeded research concerning human sexual behaviour, so that by the early 20th century scientific knowledge was largely restricted to individual case histories that had been studied by such European writers as Sigmund Freud, Havelock Ellis, and Richard, Freiherr von Krafft-Ebing. By the 1920s, however, the foundations had been laid for the more extensive statistical studies that were conducted before World War II in the United States. Of the two major organizations for sex study, one, the Institut für Sexualwissenschaft in Berlin (established in 1897), was destroyed by the Nazis in 1933. The other, the Institute for Sex Research, begun in 1938 by the American sexologist Alfred Charles Kinsey at Indiana University, Blooming-

ton, undertook the study of many aspects of human sexual behaviour. Much of the following discussion rests on the findings of the Institute for Sex Research, which comprise the most comprehensive data available. The only other country for which comprehensive data exist is Sweden.

TYPES OF BEHAVIOUR

Human sexual behaviour may conveniently be classified according to the number and gender of the participants. There is solitary behaviour involving only one individual, and there is sociosexual behaviour involving more than one person. Sociosexual behaviour is generally divided into heterosexual behaviour (male with female) and homosexual behaviour (male with male or female with female; see *Homosexuality* below.) If three or more individuals are involved it is, of course, possible to have heterosexual and homosexual activity simultaneously.

In both solitary and sociosexual behaviour there may be activities that are sufficiently unusual to warrant the label deviant behaviour. The term deviant should not be used as a moral judgment but simply as indicating that such activity is not common in a particular society. Since human societies differ in their sexual practices, what is deviant in one society may be normal in another.

Solitary behaviour. Self-masturbation is self-stimulation with the intention of causing sexual arousal and, generally, orgasm (sexual climax). Most masturbation is done in private as an end in itself but is sometimes practiced to facilitate a sociosexual relationship.

Masturbation, generally beginning at or before puberty, is very common among males, particularly young males, but becomes less frequent or is abandoned when sociosexual activity is available. Consequently, masturbation is most frequent among the unmarried. Fewer females masturbate; in the United States, roughly one-half to two-thirds have done so, as compared to nine out of ten males. Females also tend to reduce or discontinue masturbation when they develop sociosexual relationships. There is great individual variation in frequency, so that it is impractical to try to define what range could be considered "normal."

Attitudes toward masturbation

The myth persists, despite scientific proof to the contrary, that masturbation is physically harmful. Neither is there evidence that masturbation is immature behaviour; it is common among adults deprived of sociosexual opportunities. While solitary masturbation does provide pleasure and relief from the tension of sexual excitement, it does not have the same psychological gratification that interaction with another person provides; thus, extremely few people prefer masturbation to sociosexual activity. The psychological significance of masturbation lies in how the individual regards it. For some, it is laden with guilt; for others, it is a release from tension with no emotional content; and for others it is simply another source of pleasure to be enjoyed for its own sake.

The majority of males and females have fantasies of some sociosexual activity while they masturbate. The fantasy not infrequently involves idealized sexual partners and activities that the individual has not experienced and even might avoid in real life.

Since the masturbating person is in sole control of the areas that are stimulated, the degree of pressure, and the rapidity of movement, masturbation is often more effective in producing sexual arousal and orgasm than is sociosexual activity, during which the stimulation is determined to some degree by one's partner.

Orgasm in sleep evidently occurs only in humans. Its causes are not wholly known. The idea that it results from the pressure of accumulated semen is invalid because not only do nocturnal emissions sometimes occur in males on successive nights, but females experience orgasm in sleep as well. In some cases orgasm in sleep seems a compensatory phenomenon, occurring during times when the individual has been deprived of or abstains from other sexual activity. In other cases it may result from external stimuli, such as sleeping prone or having night clothing caught between one's legs. Most orgasms during sleep are accompanied by erotic dreams.

A great majority of males experience orgasm in sleep. This almost always begins and is most frequent in adolescence, tending to disappear later in life. Fewer females have orgasm in sleep, and, unlike males, they usually begin having such experience when fully adult.

Orgasm in sleep is generally infrequent, seldom exceeding a dozen times per year for males and three or four times a year for the average female.

Most sexual arousal does not lead to sexual activity with another individual. Humans are constantly exposed to sexual stimuli when seeing attractive persons and are subjected to sexual themes in advertising and the mass media. Response to such visual and other stimuli is strongest in adolescence and early adult life and usually gradually declines with advancing age. One of the necessary tasks of growing up is learning to cope with one's sexual arousal and to achieve some balance between suppression, which can be injurious, and free expression, which can lead to social difficulties. There is great variation among individuals in the strength of sex drive and responsiveness, so this necessary exercise of restraint is correspondingly difficult or easy.

Sociosexual behaviour. By far the greatest amount of sociosexual behaviour is heterosexual behaviour between only one male and one female. Heterosexual behaviour frequently begins in childhood, and, while much of it may be motivated by curiosity, such as showing or examining genitalia, many children engage in sex play because it is pleasurable. The sexual impulse and responsiveness are present in varying degrees in most children and latent in the remainder. With adolescence, sex play is superseded by dating, which is socially encouraged, and dating almost inevitably involves some physical contact resulting in sexual arousal. This contact, labelled necking or petting, is a part of the learning process and ultimately of courtship and the selection of a marriage partner.

Petting and foreplay

Petting varies from hugging, kissing, and generalized caresses of the clothed body to techniques involving genital stimulation. Petting may be done for its own sake as an expression of affection and a source of pleasure, and it may occur as a preliminary to coitus. This last form of petting is known as foreplay. In a minority of cases, but a substantial minority, petting leads to orgasm and may be a substitute for coitus. Excluding foreplay, petting is usually very stereotyped, beginning with hugging and kissing and gradually escalating to stimulation of the breasts and genitalia. In most societies petting and its escalation are initiated by the male more often than by the female, who generally rejects or accepts the male's overtures but refrains from playing a more aggressive role. Petting in some form is a near-universal human experience and is valuable not only in mate selection but as a means of learning how to interact with another person sexually.

Coitus, the insertion of the penis into the vagina, is viewed by society quite differently depending upon the marital status of the individuals. The majority of human societies permit premarital coitus, at least under certain circumstances. In more repressive societies, such as modern Western society, it is more likely to be tolerated (but not encouraged) if the individuals intend marriage. Marital coitus is usually regarded as an obligation in most societies. Extramarital coitus, particularly by wives, is generally condemned and, if permitted, is allowed only under exceptional conditions or with specified persons. Societies tend to be more lenient toward males than females regarding extramarital coitus. This double standard of morality is also seen in premarital life. Postmarital coitus (*i.e.*, coitus by separated, divorced or widowed persons) is almost always ignored. Even societies that try to confine coitus to marriage recognize the difficulty of trying to force abstinence upon sexually experienced and usually older persons.

In the United States and much of Europe, there has been, within the last century, a progressive trend toward an increase in premarital coitus. Currently in the United States, at least three-quarters of the males and over half of the females have experienced premarital coitus. The proportions for this experience vary in different groups and socioeconomic classes. In Scandinavia, the incidence of premarital coitus is far greater, exceeding the 90 percent mark in Sweden, where it is now expected behaviour.

Extramarital coitus continues to be openly condemned but is becoming more tolerated secretly, particularly if mitigating circumstances are involved. In some areas, such as southern Europe and Latin America, extramarital coitus is expected of most husbands and is accepted by society if the behaviour is not too flagrant. The wives do not generally approve but are resigned to what they believe to be a masculine propensity. In the United States, where at least half the husbands and one-quarter of the wives have extramarital coitus at some point in their lives, there have recently developed small organizations or clubs that exist to provide extramarital coitus for married couples. Despite the publicity they have engendered, however, extremely few individuals have belonged to such organizations. Most extramarital coitus is done secretly without the knowledge of the spouse. Most husbands and wives feel very possessive of their spouses and interpret extramarital activity as an aspersion on their own sexual adequacy, as indicating a loss of affection and as being a source of social disgrace.

Human beings are not inherently monogamous but have a natural desire for diversity in their sexuality as in other aspects of life. Some societies have provided a release for these desires by suspending the restraints on extramarital coitus on special occasions or with certain individuals, and in modern Western society a certain amount of extramarital flirtation or mild petting at parties is not considered unusual behaviour.

Sexuality in ceremony and religion

Discussion of sociosexual behaviour would be incomplete without some note of the role it has played in ceremony and religion. While the major religions of today are to varying degrees antisexual, many religions have incorporated sexual behaviour into their rites and ceremonies. Human beings' ancient and continuing interest in their own fertility and in that of food plants and animals makes such a connection between sex and religion inevitable, particularly among peoples with uncertain food supplies. In most religions the deities were considered to have active sexual lives and sometimes took a sexual interest in humans. In this regard it is noteworthy that in Christianity sexual behaviour is absent in heaven and sexual proclivities are ascribed only to evil supernatural beings: Satan, devils, incubi, and succubi (spirits or demons who seek out sleeping humans for sexual intercourse).

Whether or not a behaviour is interpreted by society or the individual as erotic (*i.e.*, capable of engendering sexual response) depends chiefly on the context in which the behaviour occurs. A kiss, for example, may express asexual affection (as a kiss between relatives), respect (a French officer kissing a soldier after bestowing a medal on him), reverence (kissing the hand or foot of a pope), or it may be a casual salutation and social amenity. Even something as specific as touching genitalia is not construed as sexual if done for medical reasons. In other words, the apparent motivation of the behaviour determines its interpretation.

Individuals are extremely sensitive in judging motivations: a greeting kiss, if protracted more than a second or two, takes on a sexual connotation, and recent studies show that if an adult male at a party stands closer than the length of his hand and forearm to a female, she generally imputes a sexual motive to his proximity. Nudity is construed as erotic or even as a sexual invitation—unless it occurs in a medical context, in a group consisting of but one gender, or in a nudist camp.

PHYSIOLOGICAL ASPECTS

Sexual response. Sexual response follows a pattern of sequential stages or phases when sexual activity is continued. First, there is the excitement phase marked by increase in pulse and blood pressure, an increase in blood supply to the surface of the body resulting in increased skin temperature, flushing, and swelling of all distensible body parts (particularly noticeable in the penis and female breasts), more rapid breathing, the secretion of genital fluids, vaginal expansion, and a general increase in muscle tension. These symptoms of arousal eventually increase to a near maximal physiological level, the plateau phase, which is generally of brief duration. If stimulation is continued, orgasm usually occurs. Orgasm is marked by a feeling of sudden intense pleasure, an abrupt increase in pulse rate and blood pressure, and spasms of the pelvic muscles causing vaginal contractions in the female and ejaculation by the male. Involuntary vocalization may also occur. Orgasm lasts for a few seconds (normally not over ten), after which the individual enters the resolution phase, the return to a normal or subnormal physiological state. Up to the resolution phase, males and females are the same in their response sequence, but, whereas males return to normal even if stimluation continues, continued stimulation can produce additional orgasms in females. In brief, after one orgasm a male becomes unresponsive to sexual stimulation and cannot begin to build up another excitement phase until some period of time has elapsed, but females are physically capable of repeated orgasms without the intervening "rest period" required by males.

Genetic and hormonal factors. While all normal individuals are born with the neurophysiology necessary for the sexual-response cycle described above, inheritance determines the intensity of their responses and their basic "sex drive." There is great variation in this regard: some persons have the need for frequent sexual expressions; others require very little; and some persons respond quickly and violently, while others are slower and milder in their reactions. While the genetic basis of these differences is unknown and while such variations are obscured by conditioning, there is no doubt that sexual capacities, like all other physiological capacities, are genetically determined. It is unlikely, however, that genes control the sexual orientation of normal humans in the sense of individuals being predestined to become homosexual or heterosexual. Some severe genetic abnormality can, of course, profoundly affect intelligence, sexual capacity, and physical appearance and hence the entire sexual life.

Genetic combinations in the sexes

While the normal female has 44 autosomes plus two X-chromosomes (female) and the normal male 44 autosomes plus one X-chromosome and one Y-chromosome (male), many genetic abnormalities are possible. There are females, for example, with too many X-chromosomes (44+XXX) or too few (44+X) and males with an extra female chromosome (44+XXY) or an extra male chromosome (44+XYY). No 44+YY males exist—an X-chromosome is necessary for survival, even in the womb.

One's genetic makeup determines one's hormonal status and the sensitivity of one's body to these hormones. While a disorder of any part of the endocrine system can adversely affect sexual life, the hormones most directly influencing sexuality are the androgens (male sex hormones), produced chiefly in the testicles, and the estrogens (female sex hormones), produced chiefly in the ovaries. In early embryonic life there are neither testicles nor ovaries but simply two undifferentiated organs (gonads) that can develop either into testicles or ovaries. If the embryo has a Y-chromosome, the gonads become testicles; otherwise, they become ovaries. The testicles of the fetus produce androgens, and these cause the fetus to develop male anatomy. The absence of testicles results in the development of female anatomy. Animal experiments show that, if the testicles of a male fetus are removed, the individual will develop into what seems a female (although lacking ovaries). Consequently, it has been said that humans are basically female.

After birth and until puberty, the ovaries and testicles produce comparatively few hormones, and little girls and boys are much alike in size and appearance. At puberty, however, these organs begin producing in greater abundance, with dramatic results. The androgens produced by boys cause changes in body build, greater muscular development, body and facial hair, and voice change. In girls the estrogens cause breast development, menstruation, and feminine body build. A boy castrated before puberty does not develop masculine physical characteristics and manifests in adult life more of a feminine body build, lack of masculine body and facial hair, less muscular strength, a high voice, and small genitalia. A girl who has her ovaries removed before puberty is less markedly altered but retains a childlike body build, does not develop breasts, and never menstruates. Castrated individuals or persons producing insufficient hormones can be restored to a normal condition by administration of appropriate hormones.

Beyond their role in developing the secondary sexual characteristics of the body, the hormones continue to play a role in adult life. An androgen deficiency causes a decrease in a man's sexual responsiveness, and an estrogen deficiency adversely affects a woman's fertility and causes atrophy of the genitalia. A loss of energy may also result in both men and women.

Role of androgen and estrogen

Androgen seems linked in both males and females with aggressiveness and strength of sexual drive. When androgen is given to a female in animal experiments, she becomes more aggressive and displays behaviour more typical of males—by mounting other animals, for example. Estrogen increases her sexual responsiveness and intensifies her female behaviour. Androgen given to a male often increases his sexual behaviour, but estrogen diminishes his sex drive.

In humans the picture is more complex, since human sexual behaviour and response is less dependent on hormones once adulthood has been reached. Removing androgen from an adult male reduces his sexual capacity; but this occurs gradually, and sometimes the reduction is small. Giving androgen to a normal human male generally has little or no effect since he is already producing all he can use. Giving him estrogen reduces his sex drive. Administration of androgen to an adult human female often increases her sex drive, enlarges her clitoris, and promotes the growth of facial hair. Giving estrogen to a normal woman before menopausal age generally has no effect whatsoever—probably because human females, unlike other female mammals, do not have hormonally controlled periods of "heat" (estrus).

Hormones have no connection with the sexual orientation of humans. Male homosexuals do not have more estrogens than normal males (who have a little) nor can their preferences be altered by giving them androgen.

Nervous system factors. The nervous system consists of the central nervous system and the peripheral nervous system. The brain and spinal cord constitute the central system, while the peripheral system is composed of (1) the cerebrospinal nerves that go to the spinal cord (afferent nerves), transmitting sensory stimuli and those that come from the cord (efferent nerves) transmitting impulses to activate muscles, and (2) the autonomic system, the primary function of which is the regulation and maintenance of the body processes necessary to life, such as heart rate, breathing, digestion, and temperature control. Sexual response involves the entire nervous system. The autonomic system controls the involuntary responses; the afferent cerebrospinal nerves carry the sensory messages to the brain; the efferent cerebrospinal nerves carry commands from the brain to the muscles; and the spinal cord serves as a great transmission cable. The brain itself is the coordinating and controlling centre, interpreting what sensations are to be perceived as sexual and issuing appropriate "orders" to the rest of the nervous system.

The parts of the brain thought to be most concerned with sexual response are the hypothalamus and the limbic system, but no specialized "sex centre" has been located in the human brain. Animal experiments indicate that each individual has coded in its brain two sexual response patterns, one for mounting (masculine) behaviour and one for mounted (feminine) behaviour. The mounting pattern can be elicited or intensified by male sex hormone and the mounted pattern by female sex hormone. Normally, one response pattern is dominant and the other latent but capable of being called into action when suitable circumstances occur. The degree to which such inherent patterning exists in humans is unknown.

While the brain is normally in charge, there is some reflex (*i.e.*, not brain-controlled) sexual response. Stimulation of the genital and perineal area can cause the "genital reflex": erection and ejaculation in the male, vaginal changes and lubrication in the female. This reflex is mediated by the lower spinal cord, and the brain need not be involved. Of course, the brain can override and suppress such reflex activity—as it does when an individual decides that a sexual response is socially inappropriate.

Development and change in the reproductive system. One's anatomy and sexuality change with age. The changes are rapid in intra-uterine life and around puberty but are much slower and gradual in other phases of the life cycle.

Intra-uterine sexual development

The reproductive organs first develop in the same form for both males and females: internally there are two undifferentiated gonads and two pairs of parallel ducts (Wolffian and Müllerian ducts); externally there is a genital protrusion with a groove (urethral groove) below it, the groove being flanked by two folds (urethral folds). On either side of the genital protrusion and groove are two ridgelike swellings (labioscrotal swellings). Around the fourth week of life the gonads differentiate into either testes or ovaries. If testes develop, the hormone they secrete causes the Müllerian duct to degenerate and almost vanish and causes the Wolffian duct to elaborate into the sperm-carrying tubes and related organs (the vas deferens, epididymis, and seminal vesicles, for example). If ovaries develop, the Wolffian duct deteriorates, and the Müllerian duct elaborates to form the fallopian tubes, uterus, and part of the vagina. The external genitalia simultaneously change. The genital protrusion becomes either a penis or clitoris. In the female the groove below the clitoris stays open to form the vulva, and the folds on either side of the groove become the inner lips of the vulva (the labia minora). In the male these folds grow together, converting the groove into the urethral tube of the penis. The ridgelike swellings on either side remain apart in the female and constitute the large labia (labia majora), but in the male they grow together to form the scrotal sac into which the testes subsequently descend.

At birth both male and female have all the neurophysiological equipment necessary for sexual response, although the reproductive system is not at this stage functional. Sexual interests, sexual behaviour, and sexual response are seen with increasing frequency in most children from infancy on. Even newborn males have penile erections, and babies of both sexes seem to find pleasure in genital stimulation. What appears to be orgasm has been observed in infant boys and girls, and, later in childhood, orgasm definitely can occur in masturbation or sex play.

Puberty may be defined as that short period of time (generally two years) during which the reproductive system matures and the secondary sexual characteristics appear. The ovaries and testes begin producing much larger amounts of hormones, pubic hair appears, female breasts develop, the menstrual cycle begins in females, spermatozoa and viable eggs are produced, and males experience voice change and a sudden acceleration in growth. Puberty generally occurs in females around age 12–13 and in males at about 13–14, but there is much individual variation. With puberty there is generally an intensification or the first appearance of sexual interest. Puberty marks the beginning of adolescence.

Adolescence, from a physical viewpoint, is that period between puberty and the attainment of one's maximum height. By the latter point, which occurs around age 16 in females and 18 in males, the individual has adult anatomy and physiology. In late adolescence the majority of individuals are probably at their peak in terms of sexual capacity: the ability to respond quickly and repeatedly. During this period the sex drive is at its maximum in males, although it is difficult to say whether this is also true of females, since female sexuality, in many societies, is frequently suppressed during adolescence.

Following adolescence there are about three decades of adult life during which physiological changes are slow and gradual. While muscular strength increases for a time, the changes may best be described as slow deterioration. This physical decline is not immediately evident in sexual behaviour, which often increases in quantity and quality as the individual develops more social skills and higher socio-economic status and loses some of the inhibitions and uncertainties that often impede adolescent sexuality. Indeed, in the case of the United States female, the deterioration is more than offset by her gradual loss of sexual inhibition, and the effect of age is not clear until menopausal symptoms begin. In the male, however, there is no such masking of deterioration, and the frequency of sexual activity and the intensity of interest and response slowly, but inexorably, decline.

Sexual changes in later life

If one must arbitrarily select an age to mark the beginning of old age, 50 is appropriate. By then, most females have experienced menopausal symptoms, and most males have been forced to recognize their increasing physical limitations. With menopause, the female genitalia gradually begin to atrophy and the amount of vaginal secretion diminishes—this is the direct consequence of the cessation of ovarian function and can be prevented, or the symptoms reversed, by administering estrogen. If a female has had a good sexual adjustment prior to menopause and if she does not believe in the fallacy that it spells the end of sexual life, menopause will have no adverse effect on her sexual and orgasmic ability. There is reason to believe that if a woman remains in good health and genital atrophy is prevented, she could enjoy sexual activity regardless of age. Males in good health are also capable of continuing sexual activity, although with an ever-decreasing frequency, throughout old age. The male has more difficulty in achieving erection, cannot maintain erection as long, and must have longer and longer "rest periods" between sexual acts. The amount of ejaculate becomes less, but most old males are still fertile. The Cowper's gland secretion (called "precoital mucus") diminishes or disappears entirely. According to Kinsey's data, about one-quarter of males are impotent by age 65, one-half by age 75, and three-quarters by age 80. One must remember, however, that some unknown but certainly substantial proportion of this impotence may be attributed to poor health.

In general, the female withstands the onslaughts of age better than the male. The reduction in the frequency of marital intercourse or even its abandonment is more often than not the result of male deterioration.

PSYCHOLOGICAL ASPECTS

Effects of early conditioning. Physiology sets only very broad limits on human sexuality; most of the enormous variation found among humans must be attributed to the psychological factors of learning and conditioning.

The human infant is born simply with the ability to respond sexually to tactile stimulation. It is only later and gradually that the individual learns or is conditioned to respond to other stimuli, to develop a sexual attraction to males or females or both, to interpret some stimuli as sexual and others as nonsexual, and to control in some measure his or her sexual response. In other words, the general and diffuse sexuality of the infant becomes increasingly elaborated, differentiated, and specific.

The early years of life are, therefore, of paramount importance in the development of what ultimately becomes adult sexual orientation. There appears to be a reasonably fixed sequence of development. Before age five, the child develops a sense of gender identity, thinks of himself or herself as a boy or girl, and begins to relate to others differently according to their gender. Through experience the child learns what behaviour is rewarded and what is punished and what sorts of behaviour are expected of him or her. Parents, peers, and society in general teach and condition the child about sex not so much by direct informational statements and admonitions as by indirect and often unconscious communication. The child soon learns, for example, that he can touch any part of his body or someone else's body except the anal–genital region. The child rubbing its genitals finds that this quickly attracts adult attention and admonishment or that adults will divert him or her from this activity. It becomes clear that there is something peculiar and taboo about this area of the body. This "genital taboo" is reinforced by the great concern over the child's excretory behaviour: bladder and bowel control is praised; loss of control is met by disappointment, chiding, and expressions of disgust. Obviously, the anal–genital area is not only a taboo area but a very important one as well. It is almost inevitable that the genitalia become associated with anxiety and shame. It is noteworthy that this attitude finds expression in the language of Western civilizations, as in "privates" (something to be kept hidden) and the German word for the genitals, *Scham* ("shame").

Development of genital taboo

While all children in Western civilizations experience this antisexual teaching and conditioning, a few have, in addition, atypical sexual experiences, such as witnessing or hearing sexual intercourse or having sexual contact with an older person. The effects of such atypical experiences depend upon how the child interprets them and upon the reaction of adults if the experience comes to their attention. Seeing parental coitus is harmless if the child interprets it as playful wrestling but harmful if he considers it as hostile, assaultive behaviour. Similarly, an experience with an adult may seem merely a curious and pointless game, or it may be a hideous trauma leaving lifelong psychic scars. In many cases the reaction of parents and society determines the child's interpretation of the event. What would have been a trivial and soon-forgotten act becomes traumatic if the mother cries, the father rages, and the police interrogate the child.

Some atypical developments occur through association during the formative years. A child may associate clothing, especially underclothing, stockings, and shoes with gender and sex and thereby establish the basis for later fetishism or transvestism. Others, having been spanked or otherwise punished for self-masturbation or childhood sex play, form an association between punishment, pain, and sex that could escalate later into sadism or masochism. It is not known why some children form such associations whereas others with apparently similar experience do not.

Around the age of puberty, parents and society, who more often than not refuse to recognize that children have sexual responses and capabilities, finally face the inescapable reality and consequently begin inculcating children with their attitudes and standards regarding sex. This campaign by adults is almost wholly negative—the child is told what not to do. While dating may be encouraged, no form of sexual activity is advocated or held up as model behaviour. The message usually is "be popular" (*i.e.*, sexually attractive), but abstain from sexual activity. This antisexualism is particularly intense regarding young females and is reinforced by reference to pregnancy, venereal disease, and, most importantly, social disgrace. To this list religious families add the concept of the sinfulness of premarital sexual expression. With young males the double standard of morality still prevails. The youth receives a double message, "don't do it, but we expect that you will." No such loophole in the prohibitions is offered young girls. Meanwhile, the young male's peer group is exerting a prosexual influence, and his social status is enhanced by his sexual exploits or by exaggerated reports thereof.

Effects of the double standard on the sexual relations of the young

As a result of this double standard of sexual morality, the relationship between young males and females often becomes a ritualized contest, the male attempting to escalate the sexual activity and the female resisting his efforts. Instead of mutuality and respect, one often has a struggle in which the female is viewed as a reluctant sexual object to be exploited, and the male is viewed as a seducer and aggressor who must succeed in order to maintain his self-image and his status with his peers. This sort of pathological relationship causes a lasting attitude on the part of females: men are not to be trusted; they are interested only in sex; a girl dare not smile or be friendly lest males interpret it as a sign of sexual availability, and so forth. Such an aura of suspicion, hostility, and anxiety is scarcely conducive to the development of warm, trusting relationships between males and females. Fortunately, love or infatuation usually overcomes this negativism with regard to particular males, but the average female still maintains a defensive and skeptical attitude toward men.

Western society is replete with attitudes that impede the development of a healthy attitude toward sex. The free abandon so necessary to a full sexual relationship is, in the eyes of many, an unseemly loss of self-control, and self-control is something one is urged to maintain from infancy onward. Panting, sweating, and involuntary vocalization are incompatible with the image of dignity. Worse yet is any substance once it has left the body: it immediately becomes unclean. The male and female genital fluids are generally regarded with disgust—they are not only excretions but sexual excretions. Here again, societal concern over excretion is involved, for sexual organs are also urinary passages and are in close proximity to the "dirtiest" of all places—the anus. Lastly, many individuals in soci-

ety regard menstrual fluid with disgust and abstain from sexual intercourse during the four to six days of flow. This attitude is formalized in Judaism, in which menstruating females are specifically labelled as ritually unclean.

Impediments overcome by human nature

In view of all these factors working against a healthy, rational attitude toward sex and in view of the inevitable disappointments, exploitations, and rejections that are involved in human relationships, one might wonder how anyone could reach adulthood without being seriously maladjusted. The sexual impulse, however, is sufficiently strong and persistent and repeated sexual activity gradually erodes the inhibitions and any sense of guilt or shame. Further, all humans have a deep need to be esteemed, wanted, and loved. Sexual activity with another is seen as proof that one is attractive, desired, valued, and possibly loved—a proof very necessary to self-esteem and happiness. Hence, even among the very inhibited or those with weak sex drive, there is this powerful motivation to engage in sociosexual activity.

Most persons ultimately achieve at least a tolerable sexual adjustment. Some unfortunates, nevertheless, remain permanently handicapped, and very few completely escape the effects of society's antisexual conditioning. While certain inhibitions and restraints are socially and psychologically useful—such as deferring gratification until circumstances are appropriate and modifying behaviour out of regard for the feelings of others—most people labour under an additional burden of useless and deleterious attitudes and restrictions.

Sexual problems. Sexual problems may be classified as physiological, psychological, and social in origin. Any given problem may involve all three categories; a physiological problem, for example, will produce psychological effects, and these may result in some social maladjustment.

Physiological problems of a specifically sexual nature are rather few. Only a small minority of people suffer from diseases of or deficient development of the genitalia or that part of the neurophysiology governing sexual response. Many people, however, experience at some time sexual problems that are by-products of other pathologies or injuries.

Vaginal infections, for example, retroverted uteri, prostatitis, adrenal tumours, diabetes, senile changes of the vagina, and cardiovascular conditions may cause disturbance of the sexual life. In brief, anything that seriously interferes with normal bodily functioning generally causes some degree of sexual trouble. Fortunately, the great majority of physiological sexual problems are solved through medication or surgery. Generally, only those problems involving damage to the nervous system defy therapy.

Psychological problems constitute by far the largest category. They are not only the product of socially induced inhibitions, maladaptive attitudes, and ignorance but also of sexual myths held by society. An example of the latter is the idea that good, mature sex must involve rapid erection, protracted coitus, and simultaneous orgasm. Magazines, marriage books, and general sexual folklore reinforce these demanding ideals, which cannot always be met and hence give rise to anxiety, guilt, and feelings of inadequacy.

Premature ejaculation and its therapy

Premature ejaculation is a common problem, especially for young males. Sometimes this is not the consequence of any psychological problem but the natural result of excessive tension in a male who has been sexually deprived. In such cases, more frequent coitus solves the problem. Premature ejaculation is difficult to define. The best definition is that offered by the American sexologists, William Howell Masters and Virginia Eshelman Johnson, who say that a male suffers from premature ejaculation if he cannot delay ejaculation long enough to induce orgasm in a sexually normal female at least half the time. This generally means that vaginal penetration with some movement (although not continuous) must be maintained for more than one minute. The average American male ejaculates in two or three minutes after vaginal penetration, a coital duration sufficient to cause orgasm in most females the majority of the time. Various methods of preventing premature ejaculation have been tried. One is for the male to excite the female more during the foreplay so that she reaches orgasm more rapidly after penetration, but this technique often excites the male as well and defeats its purpose. Another common method is for the male to think of nonsexual matters, which may prove effective but reduces his pleasure. The most effective therapy is that advocated by Masters and Johnson in which the female brings the male nearly to orgasm and then prevents the male's orgasm by briefly compressing the penis between her fingers just below the head of the penis. The couple come to realize that premature ejaculation can thus be easily prevented, their anxiety disappears, and ultimately they can achieve normal coitus without resorting to this squeeze technique.

Impotence

Erectile impotence is almost always of psychological origin in males under 40; in older males physical causes are more often involved. Fear of being impotent frequently causes impotence, and, in many cases, the afflicted male is simply caught up in a self-perpetuating problem that can be solved only by achieving a successful act of coitus. In other cases, the impotence may be the result of disinterest in the sexual partner, fatigue, distraction because of nonsexual worries, intoxication, or other causes—such occasional impotency is common and requires no therapy.

Some males, however, are chronically impotent and require psychotherapy or behaviour therapy. Such impotency is thought to be the result of deep-seated causal factors such as unconscious feelings of hostility, fear, inadequacy, or guilt. Primary impotence, the inability to ever have achieved erection sufficient for coitus, is more difficult to treat than the far more common secondary impotence, which is impotence in a male who was formerly potent.

Ejaculatory impotence, the inability to ejaculate in coitus, is quite rare and is almost always of psychogenic origin. It seems associated with ideas of contamination or with memories of traumatic experiences. Occasional ejaculatory inability may be expected in older men or in any male who has exceeded his sexual capacity.

Vaginismus and dyspareunia

Vaginismus is a powerful spasm of the pelvic musculature constricting the vagina so that penetration is painful or impossible. It seems wholly due to antisexual conditioning or psychological trauma and serves as an unconscious defense against coitus. It is treated by psychotherapy and by gradually dilating the vagina with increasingly large cylinders.

Dyspareunia, painful coitus, is generally physical rather than psychological. It is mentioned here only because some inexperienced females fear they cannot accommodate a penis without being painfully stretched. This is a needless fear since the vagina is not only highly elastic but enlarges with sexual arousal, so that even a small female can, if aroused, easily receive an exceptionally large penis.

Disparity in sexual desire constitutes the most common sexual problem. It is to some extent inescapable, since differences in the strength of the sexual impulse and the ability to respond are based on neurophysiological differences. Much disparity, however, is the result of inhibition or of one person having been subjected to more sexual stimuli during the day than the other. The partner who has been seeing sexually attractive persons periodically during the day and who may have had an opportunity to relax on the way back from the office or store is naturally more interested in coitus than the partner who has not been exposed to sexual stimuli. Another cause of disparity is a difference in viewpoint. Perhaps one person anticipates coitus as a palliative to compensate for the trials and tribulations of life, whereas another may be interested in sex only if the preceding hours have been reasonably problem-free and happy. Even in cases of neurophysiological differences in sex drive, the less-motivated partner can be trained to a higher level of interest, since most humans operate well below their sexual capacities.

Psychological fatigue, a growing disinterest in sexual behaviour with a particular partner, sometimes constitutes a problem. Humans are subject to monotony, and coitus may become routine or even a chore. Lessening frequencies of marital coitus are more often the result of this than of age. The solution lies in varying the time, the setting, and in breaking away from habitual techniques and positions.

Preferences for or antipathies toward particular positions, techniques, or times frequently cause trouble. One partner may desire mouth-genital contact or anal stimulation that the other partner finds disagreeable or perverse. Some wish to have coitus in the light, others insist upon darkness; some prefer morning, others evening. The possibilities for disagreement are legion. Even if disagreements stemming from needless inhibition are overcome, there still remain disparities in preference, and these should be met by the philosophy that, by giving pleasure to another, one obtains pleasure. Needless to say, no partner should insist upon that which is abhorrent to the other after the latter has made honest attempts to cooperate.

Anorgasmy and frigidity

Lack of female orgasm, anorgasmy, is a very frequent problem. One should differentiate between females who become sexually aroused but do not reach orgasm and those who do not become aroused. Only the latter merit the label frigid. It is common for females not to achieve orgasm during the first weeks or months of coital activity. It is almost as though many females must learn how to have orgasm, for after having had one they respond with increasing frequency. In some cases, the female initially has no idea how to copulate effectively and simply lies passive, expecting the male to bring her to orgasm. Other females resist orgasm because the feeling of being swept away and losing control is frightening. In most cases, however, anorgasmy is simply the result of years of inhibition—having been trained since childhood to avoid yielding to the sexual impulse, it is difficult to metamorphose into a responsive and orgasmic being. In the final analysis, anorgasmy is psychological in origin; few, if any, females lack the neurophysiology necessary for orgasm, and anthropology shows that in sexually permissive societies virtually all females have little difficulty in attaining orgasm in coitus.

Anorgasmy is treated by removing inhibitions, by teaching coital techniques, and by inducing orgasm through noncoital methods. The effective therapist should also impress upon the female that not reaching orgasm is no sign of failure or inadequacy on her part or her partner's and that sexual activity is very pleasurable to both, even if orgasm does not ensue. Indeed, some females derive great pleasure and satisfaction without orgasm, a fact that should be made known to anxious male partners. Too great a concern over orgasm defeats itself. As Kinsey once pointed out, thinking is the enemy of sexual pleasure, and a female can scarcely have orgasm if she is worrying about whether she will attain it or not and if she senses that her partner is mentally turning the pages of a marriage manual.

Lastly, sexual problems are often perpetuated by the inability of the partners to communicate freely their feelings to one another. There is a curious and unfortunate reticence about informing one's partner as to what does or does not contribute to one's pleasure. The partner must function on a trial-and-error basis, ever on the alert for signs indicating the efficacy of his or her efforts. This muteness is even more pronounced when it comes to an individual making suggestions to the partner. Many persons feel that a suggestion or request would be interpreted by the partner that he or she had been inept or at least remiss. As with any other problems, sexual problems can be overcome or ameliorated only if the individuals concerned communicate effectively.

SOCIAL AND CULTURAL ASPECTS

The effects of societal value systems on human sexuality are, as has already been mentioned, profound. The American anthropologist George P. Murdock summarized the situation, saying:

> All societies have faced the problem of reconciling the need of controlling sex with that of giving it adequate expression, and all have solved it by some combination of cultural taboos, permissions, and injunctions. Prohibitory regulations curb the socially more disruptive forms of sexual competition. Permissive regulations allow at least the minimum impulse gratification required for individual well-being. Very commonly, moreover, sex behavior is specifically enjoined by obligatory regulations where it appears directly to subserve the interests of society.

The historical heritage is, of course, the foundation upon which the current situation rests. Western civilizations are basically Greco-Roman in social organization, philosophy, and law, with a powerful admixture of Judaism and Christianity. This historical mixture contained incompatible elements: individual freedom was cherished, yet there was a great emphasis on law and proper procedure; the pantheism of the Greeks and Romans clashed with Judeo-Christian monotheism; and the sexual permissiveness of Hellenistic times was answered by the antisexuality of early Christianity.

Effects of Christianity on sexual attitudes

In terms of sex, the most important factor was Christianity. While other vital aspects of human life, such as government, property rights, kinship, and economics, were influenced to varying degrees, sexuality was singled out as falling almost entirely within the domain of religion. This development arose from an ascetic concept shared by a number of religions, the concept of the good spiritual world as opposed to the carnal materialistic world, the struggle between the spirit and the flesh. Since sex epitomizes the flesh, it was obviously the enemy of the spirit. Beginning in the 2nd century, Western Christianity was heavily influenced by this dichotomous philosophy of the Gnostics; sex in any form outside of marriage was an unmitigated evil and, within marriage, an unfortunate necessity for purposes of procreation rather than pleasure. The powerful antisexuality of the early Christians (note that neither God nor Christ has a wife and that marriage does not exist in heaven) was in part due to their apocalyptic vision of life: they anticipated that the end of the world and the Last Judgment would soon be upon them. There was no time for a gradual weaning away from the flesh; an immediate and drastic approach was necessary. Indeed, such excessive antisexuality developed that the church itself was finally moved to curb some of its more extreme forms.

As it became evident that human existence was going to continue for some unforeseeable length of time and as occasional intelligent theologians made themselves felt, antisexuality was ameliorated to some extent but still remained a foundation stone of Christianity for centuries. This attitude was particularly unfortunate for women, to whom most of the sexual guilt was assigned. Women, like the original temptress Eve, continued to attract men to commit sin. They were spiritually weak creatures prone to yield to carnal impulses. This is, of course, a classic example of projecting one's own guilty desires upon someone else.

Ultimately, legal control over sexual behaviour passed from the church to the state, but in most instances the latter simply perpetuated the attitudes of the former. Priests and clergymen frequently continued to exert powerful extralegal control: denunciations from the pulpit can be as effective as statute law in some cases. Although religion has weakened as a social control mechanism, even today liberalization of sex laws and relaxation of censorship have often been successfully opposed by religious leaders. On the whole, however, Christianity has become progressively more permissive, and sexuality has come to be viewed not as sin but as a God-given capacity to be used constructively.

Apart from religion, the state sometimes imposes restrictions for purely secular reasons. The more totalitarian a government, the more likely it is to restrict or direct sexual behaviour. In some instances, this comes about simply as the consequence of a powerful individual (or individuals) being in a position to impose ideas upon the public. In other instances, one cannot escape the impression that sex, being a highly personal and individualistic matter, is recognized as antithetical to the whole idea of strict governmental control and supervision of the individual. This may help explain the rigid censorship exerted by most totalitarian regimes over sexual expression. It is as though such a government, being obsessed with power, cannot tolerate the power the sexual impulse exerts on the population.

Social control of sexual behaviour. Societies differ remarkably in what they consider socially desirable and undesirable in terms of sexual behaviour and consequently

differ in what they attempt to prevent or promote. There appear, however, to be four basic sexual controls in the majority of human societies. First, to control endless competition, some form of marriage is necessary. This not only removes both partners from the competitive arena of courtship and assures each of a sexual partner, but it allows them to devote more time and energy to other necessary and useful tasks of life. Despite the beliefs of earlier writers, marriage is not necessary for the care of the young; this can be accomplished in other ways.

Four major areas of sexual control

Second, control of forced sexual relationships is necessary to prevent anger, feuding, and other disruptive retribution.

Third, all societies exert control over whom one is eligible to marry or have as a sexual partner. Endogamy, holding the choice within one's group, increases group solidarity but tends to isolate the group and limit its political strength. Exogamy, forcing the individual to marry outside the group, dilutes group loyalty but increases group size and power through new external liaisons. Some combination of endogamy and exogamy is found in most societies. All have incest prohibitions. These are not based on genetic knowledge. Indeed, many incest taboos involve persons not genetically related (father–stepdaughter, for example). The prime reason for incest prohibition seems to be the necessity for preventing society from becoming snarled in its own web: every person has a complex set of duties, rights, obligations, and statuses with regard to other people, and these would become intolerably complicated or even contradictory if incest were freely permitted.

Fourth, there is control through the establishment of some safety-valve system: the formulation of exceptions to the prevailing sexual restrictions. There is the recognition that humans cannot perpetually conform to the social code and that well-defined exceptions must be made. There are three sorts of exceptions to sexual restrictions: (1) Divorce: while all societies encourage marriage, all realize that it is in the interest of society and the individual to terminate marriage under certain conditions. (2) Exceptions based on kinship: many societies permit or encourage sexual activity with certain kin, even after marriage. Most often these kin are a brother's wife or a wife's sister. In addition, sexual "joking relationships" are often expected between brothers-in-law, sisters-in-law, and cousins. While coitus is not involved, there is much explicit sexual banter, teasing, and humorous insult. (3) Exceptions based on special occasions, ranging from sexual activity as a part of religious rites to purely secular ceremonies and celebrations wherein the customary sexual restrictions are temporarily lifted.

Turning to particular forms of sexual behaviour, one learns from anthropology and history that extreme diversity in social attitude is common. Most societies are unconcerned over self-masturbation since it does not entail procreation or the establishment of social bonds, but a few regard it with disapprobation. Sexual dreams cause concern only if they are thought to be the result of the nocturnal visitation of some spirit. Such dreams were once attributed to spirits or demons known as incubi and succubi, who sought out sleeping humans for sexual intercourse.

Petting among most preliterate societies is done only as a prelude to coitus—as foreplay—rather than as an end in itself. In some parts of sub-Saharan Africa, however, petting is used as a premarital substitute for coitus in order to preserve virginity and avoid pregnancy. There is great variation in petting and foreplay techniques. Kissing is by no means universal, as some groups view the mouth as a biting and chewing orifice ill-suited for expressing affection. While some societies emphasize the erotic role of the female breast, others—such as the Chinese—pay little attention to it. Still others regard oral stimulation of the breast unseemly, being too akin to infantile suckling. Although manual stimulation of the genitalia is nearly universal, a few peoples abstain because of revulsion toward genital secretions. Not much information exists on mouth–genital contact, and one can say only that it is common among some peoples and rare among others.

A considerable number of societies manifest scratching and biting in conjunction with sexual activity, and most of this is done by the female. Sadomasochism in any other form, however, is conspicuous by its absence in preliterate societies.

An enumeration of the societies that permit or forbid premarital coitus is complicated not only by the double standard but also by the fact that such prohibition or permission is often qualified. As a rough estimate, however, 40 to 50 percent of preliterate or ancient societies allowed premarital coitus under certain conditions to both males and females. If one were to count as permissive those groups that theoretically disapprove but actually condone such coitus, the percentage would rise to perhaps 70.

Frequency of marital coitus

In marital coitus, when sexual access is not only permitted but encouraged, one would expect considerable uniformity in frequency of coitus. This expectation is not fulfilled: social conditioning profoundly affects even marital coitus. On one Irish island reported upon by a researcher, for example, marital coitus is best measured in terms of per year, and among the Cayapas of Ecuador, a frequency of twice a week is something to boast of. The coital frequencies of other groups, on the other hand, are nearer to human potential. In one Polynesian group, the usual frequency of marital coitus among individuals in their late 20s was 10 to 12 per week, and in their late 40s the frequency had fallen to three to four. The African Bala, according to one researcher, had coitus on the average of once or twice per day from young adulthood into the sixth decade of life.

Marital coitus is not unrestricted. Coitus during menstruation or after a certain stage of pregnancy is generally taboo. After childbirth a lengthy period of time must often elapse before coitus can resume, and some peoples abstain for magical reasons before or during warfare, hunting expeditions, and certain other important events or ceremonies. In modern Western society one finds menstrual, pregnancy, and postpartum taboos perpetuated under an aesthetic or medical guise, and coaches still attempt to force celibacy upon athletes prior to competition.

Extramarital coitus provides a striking example of the double standard: it is expected, or tolerated, in males and generally prohibited for females. Very few societies allow wives sexual freedom. Extramarital coitus with the husband's consent, however, is another matter. Somewhere between two-fifths and three-fifths of preliterate societies permit wife lending or allow the wife to have coitus with certain relatives (generally brothers-in-law) or permit her freedom on special ceremonial occasions. The main concern of preliterate societies is not one of morality, but of more practical considerations: does the act weaken kinship ties and loyalty? Will it damage the husband's social prestige? Will it cause pregnancy and complicate inheritance or cause the wife to neglect her duties and obligations? Most foreign of all to Western thinking is that of those peoples whose marriage ceremony involves the bride having coitus with someone other than the groom, yet it is to be recalled that this practice existed to a limited extent in medieval Europe as jus primae noctis, the right of the lord to the bride of one of his subjects.

Sexual deviations and sex offenses are, of course, social definitions rather than natural phenomena. What is normative behaviour in one society may be a deviation or crime in another. One can go through the literature and discover that virtually any sexual act, even child–adult relations or necrophilia, has somewhere at some time been acceptable behaviour. Homosexuality is permitted in perhaps two-thirds of human societies. In some groups it is normative behaviour, whereas in others it is not only absent but beyond imagination. Generally, it is not an activity involving most of the population but exists as an alternative way of life for certain individuals. These special individuals are sometimes transvestites—that is, they dress and behave like the opposite sex. Sometimes they are regarded as curiosities or ridiculed, but more often they are accorded respect and magical powers are attributed to them. It is noteworthy, however, that aside from these transvestites, exclusive homosexuality is quite rare in preliterate societies.

In conclusion, the cardinal lesson of anthropology is that no type of sexual behaviour or attitude has a universal, inherent social or psychological value for good or evil—the

whole meaning and value of any expression of sexuality is determined by the social context within which it occurs.

Class distinctions. Differences in sexual behaviour between classes within technologically developed societies are very marked. Civilizations are made up of class hierarchies, and the different subgroups normally develop their own value systems. Most of the knowledge of the sexual behaviour and attitudes of ancient cultures is that of the upper or ruling class; the behaviour and feelings of the slaves and peasants were seldom recorded. There is the impression—probably a correct one—that throughout history the lower socio-economic class was the most permissive. Sex has always been one of the few pleasures of the poor and oppressed. On the other hand, one must not overlook the fact that a fanatical Puritanism can also flourish at the bottom of the social scale, and, hence, one can never assume that low status and sexual permissiveness are inevitably linked.

The Kinsey studies showed considerable social class differences in sexuality in the United States, chiefly in that the lower class was more tolerant of nonmarital coitus. More recent studies indicate that these class differences have rapidly broken down. Increased literacy and the influence of mass media have made the population more homogeneous in sexual attitudes. One can find, moreover, reversals of the previous pattern: a lower class person on the way up the social ladder may be quite conservative in his sexual views, feeling that this facilitates upward mobility, whereas the person secure in his or her high social status often feels that he or she can afford to flout convention. Actually, the most sexually liberal are those at the very bottom, who have nothing to lose, and those at the very top, who are beyond social retribution.

Breakdown of class differences in sexual behaviour

The great middle class remains the bastion of traditionalism, and it is here that the double standard of morality is most prominent. The intellectualized liberalism of the upper level seeps down only slowly, and the pragmatic egalitarianism of the lower level does not penetrate far upward.

Economic influences. Systems of production and distribution have had a growing influence on sexual behaviour since the Industrial Revolution. The old family pattern was inexorably disrupted by the rise of the industrial state. Children were no longer kept at home to share in the work and be economic assets but left for school or for nonfamily employment, and the degree of parental control diminished. The "working wife" employed outside the home, once found only among the impoverished, has gradually become the typical wife. With her enhanced economic power and her greater association with people outside the home, she became less a chattel. As the population left the family farm and tight-knit small communities for anonymous big-city existence, not only parental but societal controls over behaviour were weakened. Society became increasingly nomadic with improved transportation and job opportunities. Cultural and ethnic subgroups that formerly would have had little contact were thrown together in the same schools, factories, offices, and neighbourhoods.

All of this vast uprooting and rearranging naturally altered sexual attitudes and behaviour. The individual no longer had the option of choosing to conform or depart from a rather clear-cut sexual moral code but instead was faced with a multiplicity of choices of varying degrees of social acceptability. The major sexual change—one still in progress—was the emancipation of women, which brought with it an increasing acceptance of premarital sexual activity, the concept of woman as a human being with her own sexual needs and rights, and the possibility of terminating an unhappy marriage without incurring serious social censure. A second major change was the erosion of simplistic value systems: with increased mobility and social mixing, the individual learned that the values and attitudes he or she had unquestioningly accepted were not necessarily shared by neighbours and co-workers. As a result, life became not only more complex but more permissive. This growing tolerance has in recent decades extended, to a limited extent, to homosexuality. There is no evidence that homosexuality or other deviant behaviour has measurably increased as a result of society's urbanization and technological progress, but one gains the impression of an increase simply because these topics, previously unmentionable, are now openly discussed in the mass media.

While the old monolithic value systems broke down and individuals were accorded a wider variety of choices in terms of sexual life, there developed a paradoxical trend toward homogeneity as a result of mobility, the mass media, and increasing economic parity. Geographical and social-class differences in sexual attitudes and behaviour have steadily lessened. The plumber's family and the banker's family are now indistinguishable in terms of dress; both have automobiles; their offspring attend the same schools; and they share the same newspapers, magazines, and television programs. One might summarize by saying that society is homogeneous in that everyone now has available a wide diversity of sexual attitudes and activities.

Legal regulation. Sex laws, the origins of which, as mentioned above, are found within the church, are unique in one important respect. Whereas all other laws are basically concerned with the protection of person or property, the majority of sex laws are concerned solely with maintaining morality. The issue of morality is minimal in other laws: one can legitimately evict an impoverished old couple from their mortgaged home or sentence a hungry man for stealing food. Only in the realm of sex is there a consistent body of law upholding morality.

Early examples of sex laws

The earliest sex laws of which there is knowledge are from the Near East and date back to the 2nd millennium BC. They are remarkable in three respects: there are great omissions—certain acts are not mentioned whereas others receive detailed attention; some laws seem almost contradictory; and penalties are often extraordinarily severe. One obtains the distinct impression that these laws were case law—that is, laws formulated upon specific cases as they arose rather than being the result of lengthy judicial deliberation done in advance. These laws influenced Judaic and, hence, Christian thinking, and some were immortalized in the Bible, chiefly in Leviticus.

As mentioned earlier, when secular law replaced religious law, there was rather little change in content. In Europe the Napoleonic Code represented a break with tradition and introduced some measure of sexual tolerance, but in England and the United States there was no such rift with the past. In the latter country, as each new state joined the union, its sex laws simply duplicated, to a great extent, those of pre-existing states; legislators were disinclined to debate sexual issues or to risk losing votes by discarding or weakening sex laws.

Sex laws may be grouped in three categories: (1) Those concerned with protection of person. These are based on the element of consent. These otherwise logical laws become problematic when society deems that minors, mental retardates, and the insane are incapable of giving consent—hence, coitus with them is rape. (2) Those concerned with preventing offense to public sensibilities. Statutes preclude public sexual activity, exhibitionism, and offensive solicitation. (3) Those concerned with maintaining sexual morality. These constitute the majority of sex laws, covering such items as premarital coitus, extramarital coitus, incest, homosexuality, prostitution, peeping, nudity, animal contact, transvestism, censorship, and even specific sexual techniques—chiefly oral or anal. Laws relating to sexual conduct and morality are generally far more extensive in the United States than in western Europe and most other areas of the world.

In recent years, in Europe and the United States, a number of highly respected legal, medical, and religious organizations have deliberated on the issue of the legal control of human sexuality. They have been unanimous in the conclusion that, while laws protecting person and public sensibilities should be retained, the purely moral laws should be dropped. What consenting adults do in private, it is argued, should not be subject to legal control.

In the final analysis, sexuality, like any other vital aspect of human life, must be dealt with on an individual or societal level with a combination of rationality, sensitivity, and tolerance if society is to avoid personal and social problems arising from ignorance and misconception.

(P.H.Ge./Ed.)

SEXUALLY TRANSMITTED DISEASES

Infections transmitted primarily by sexual contact are referred to as sexually transmitted diseases (STDs). Caused by a variety of microbial agents that thrive in warm, moist environments such as the mucous membranes of the vagina, urethra, anus, and mouth, STDs are diagnosed most frequently in individuals who engage in sexual activity with many partners.

In the past, a disease transmitted sexually was more commonly called a venereal disease, or VD, a name that was applied to only a few infections such as gonorrhea and syphilis. Actually more than 20 STDs have been identified, and infections caused by *Chlamydia trachomatis,* herpes simplex virus, and human papillomavirus, although underreported, are believed to be more prevalent than gonorrhea in the United States. Although the incidence of some STDs has reached epidemic proportions, it was not until the advent of the acquired immunodeficiency syndrome (AIDS) that the need to restrain the transmission of these diseases gained serious attention.

AIDS is a deadly disease for which there is no known cure. This fact has made prevention of the spread of HIV (see below) infection a top priority of the health-care community, with education concerning safer sexual practices at the fore. The "safe sex" strategy, which includes encouraging the use of condoms or the practice of abstinence, has been introduced to prevent the spread not only of AIDS but of all STDs. Stemming the transmission of disease rather than relying on treatment is the basic tenet of the safe sex doctrine.

Preventing the transmission of STDs is also important because many of these diseases do not produce initial symptoms of any significance. Thus, they often go untreated, increasing their spread and the incidence of serious complications; untreated chlamydial infections in women are the primary preventable cause of female sterility.

Common sexually transmitted organisms. Bacteria, parasites, and viruses are the most common microbial agents involved in the sexual transmission of disease. Bacterial agents include *Neisseria gonorrhoeae,* which causes gonorrhea and predominantly involves the ureter in men and the cervix in women, and *Treponema pallidum,* which is responsible for syphilis. The parasite *Chlamydia trachomatis* causes a variety of disorders—in women, pelvic inflammatory disease (inflammation of the reproductive organs) and, in men, epididymitis. Sexually transmitted viral agents include the human papillomavirus, which causes genital warts. Infection by this virus, of which there are more than 100 types, is the major cause of cervical carcinoma. Herpes simplex virus II is the usual causative agent of genital herpes, a condition in which ulcerative blisters form on the mucous membranes of the genitalia.

Acquired immunodeficiency syndrome. AIDS is caused by the human immunodeficiency virus (HIV), a pernicious infectious agent that attacks the immune system, leading to its progressive destruction. The virus is found in highest concentrations in the blood, semen, and vaginal and cervical fluids of the human body and can be harboured asymptomatically for 10 years or more. Although the primary route of transmission is sexual, HIV also is spread by the use of infected needles among intravenous drug users, by the exchange of infected blood products, and from an infected mother to her fetus during pregnancy.

The progression of the syndrome does not follow a defined path; instead nonspecific symptoms reflect the myriad effects of a failing immune system. These symptoms are referred to as AIDS-related complex (ARC) and include fever, rashes, weight loss, and wasting. Opportunistic infections such as *Pneumocystis carinii* pneumonia, neoplasms such as Kaposi's sarcoma, and central nervous system dysfunction are also common complications. The patient eventually dies, unable to mount an immunologic defense against the constant onslaught of infections.

A blood test can be used to detect HIV infection before the symptoms begin to manifest themselves, and all individuals who may be at even the slightest risk of infection are encouraged to be tested in order to prevent the unknowing spread of HIV to others. Identification of infection before the onset of the disease, however, does not promise a better prognosis; the vast majority of those infected with HIV will ultimately succumb to AIDS. Although development of a vaccine is being pursued, it is not yet available and education remains the best way to prevent transmission of this lethal disease. (Ed.)

HOMOSEXUALITY

The term *homosexual* is used to characterize individuals who prefer romantic attachments and sexual relations with persons of the same sex. While the term is used for both sexes, a female homosexual is often referred to as a lesbian, while a male homosexual is often referred to as gay. Homosexual behaviour consists of having a romantic relationship and sexual interaction with a partner of the same sex.

Homosexual behaviour is controversial if not legally proscribed in some societies and has often been subject to prejudice. In its extreme form, this antipathy is known as heterosexism, which is defined by psychologist Gregory M. Herek "as an ideological system that denies, denigrates, and stigmatizes any nonheterosexual form of behavior, identity, relationship, or community." The word *homophobia* refers to a fear or hatred of homosexuals or homosexuality. As with the objects of most prejudice, homosexuals are seen in terms of various stereotypes. Gay men are thought of as effeminate, lesbians as masculine. "Characteristic" postures and behaviours are attributed to both. For example, some associate gay men with promiscuity. Scientific and social science research, however, has confirmed what to some may be obvious—that stereotypes of homosexuals are inaccurate and indeed can contribute to much of the difficulty that some homosexuals experience.

Classification and prevalence. Because of the difficulty in classification, sex researcher Alfred C. Kinsey and his associates at the Institute for Sex Research (now the Kinsey Institute at Indiana University) created in 1948 an influential, but controversial, seven-point scale that included exclusively heterosexual behaviour at one extreme and exclusively homosexual behaviour at the other. For this study, the critical factors in classifying an individual as homosexual, heterosexual, or bisexual were determined to be: (1) which sex or sexes one desires for the formation of romantic bonds and (2) how easily one responds sexually to the chosen partners.

Although Kinsey's sources and sampling methods have been widely debated, his studies represent some of the earliest research on homosexuality. According to the extensive data gathered by Kinsey and his associates for his landmark work *Sexual Behavior in the Human Male* (1948), a study of 5,300 males in the United States, approximately 50 percent had a same-sex genital experience before puberty. Twenty-five percent had more than incidental homosexual experience for at least three years between the ages of 16 and 55 years, and 37 percent had at least one homosexual experience leading to orgasm after puberty. Ten percent were exclusively homosexual for a period of at least three years between the ages of 16 and 55.

By age 30, one-quarter of the 5,940 U.S. females interviewed by Kinsey (*Sexual Behavior in the Human Female,* 1953) had recognized erotic responses to the same sex. About 20 percent reported some same-sex erotic experience; but, because these figures included "casual" contacts that elicited erotic response, the more insightful statistic was represented by same-sex experience to orgasm: 13 percent of women had a homosexual experience to orgasm prior to age 45. Approximately 2 to 3 percent reported having exclusive homosexual experience.

This range is consistent with more recent, cross-cultural studies that show the worldwide incidence of exclusive homosexuality to be 2 to 10 percent. It is likely, therefore, that, although sexual liberation movements of the late 20th century brought greater openness to sexuality in general, (including homosexuality), there was not a measurable increase in the number of individuals having homosexual experiences. What increased significantly was the candour with which sexual orientation could be discussed and displayed. Over time, a new sense of openness toward homosexual inclination was seen in a number of governments,

businesses, schools, and other institutions. For example, churches and synagogues, especially in urbanized areas of the United States and western Europe, initiated discussions of the Bible and homosexuality, while schools and universities offered courses on "queer" literature or in gay and lesbian studies.

Prenatal and environmental influences. The "biologic versus learned" dichotomy regarding the etiology of homosexual orientation—whether it is determined by inborn factors or by environmental influences—continues to muddle scientific investigation of the interaction of prenatal and postnatal factors that contribute to all sexual orientation. There is evidence to suggest that human beings are born with a sexual potential and that heterosexual, homosexual, bisexual, or asexual preferences unfold during the experiences of childhood and adolescence. This is not to say, however, that prenatal and genetic factors are unimportant. There is also evidence that suggests possible predispositions to sexual orientation. Evidence that humans are in some cases born bipotential (that is, having the potential to develop as either of two ways) with regard to sexual orientation comes from several sources. The first is the study of clinical cases of intersexuality—that is, individuals born with ambiguous genitalia (a condition known as hermaphroditism) and raised in the opposite gender of their genetic, or chromosomal, sex. (Males have an X and a Y sex chromosome; females have two X sex chromosomes. The combination that an individual possesses can be determined by testing.) Medical psychologist John W. Money and his associates at Johns Hopkins University studied more than 50 individuals who were raised in the sex opposite of their genetic sex. In addition, their study covered an unusual case of chromosomal male identical twins who were reared as opposite sexes. In each of the cases of intersexuality, as well as the twins' case, the gender identity and gender role (that is, the individual's personal sense of masculinity or femininity and behavioral manifestations) that developed were compatible with the sex of rearing regardless of the genetic and prenatal classification. At puberty these individuals were generally attracted to the sex opposite of their sex of rearing. The conclusions of these studies suggested that sexual orientation is primarily established in postnatal experiences.

Anthropologic research provides further information on the influences of nature and nurture. In almost two-thirds of the 76 societies that anthropologist Clellan S. Ford and psychobiologist Frank A. Beach reviewed in *Patterns of Sexual Behavior* (1951), homosexual activities were considered acceptable under certain circumstances. For example, homosexual activity might be ritualized during childhood and adolescence. In none of these societies was there a record of exclusive homosexuality; rather, heterosexual pair-bonding occurred during adulthood, and homosexual activity either stopped in adulthood or was permitted under special circumstances. So it can be seen that exclusive homosexuality—and exclusive heterosexuality—are atypical in many societies. Polarizations to homosexuality and to heterosexuality are thought by many to be a product of individual cultures.

Programming of erotic orientation. The programming, or conditioning, of erotic orientation develops as part of the teaching of gender identity and gender roles to children. In traditional families, children learn to imitate the parent of the same sex and complement the opposite-sexed parent. In nontraditional families, such as those headed by a same-sex couple, research has indicated that children are no more or less likely to become gay or lesbian than are children from traditional families. The same is true for children parented by a single parent. A child can reliably distinguish the sex of other people by age five. By this point boys, for example, are aware of the sexual anatomy of both sexes as well as the stereotypic characteristics of the masculine sex role, and they are also aware that they are, in most cases, expected to marry a woman in adulthood. Developmental cues such as this suggest the acculturative aspect of heterosexuality.

Moreover, the programming that encourages heterosexuality and discourages homosexuality, even close physical contact with the same sex, generally results in the stunting of any homosexual fantasy or behaviour in most prepubescents. Heterosexuality is left to unfold.

Even though homosexual behaviour is not sanctioned in many cultures, homosexual erotic fantasies may persist. Study results published in the 1970s by sex researchers Virginia Johnson and William Masters and by author Nancy Friday suggested that many heterosexuals use homosexual imagery to enhance their sexual arousal but do not actually have an interest in having a homosexual experience. Similarly, many homosexuals were found to use heterosexual imagery to enhance their arousal but did not desire heterosexual experiences.

Why some individuals develop homosexual orientations despite contrary socialization is not yet completely known. There are obviously prenatal and postnatal contributing factors that interact to cause this particular sexual orientation. Still—even in the face of lingering social intolerance—the equality granted gays and lesbians by a growing number of governments and businesses, along with the presence of a growing body of literary, musical, and visual works by and about gays and lesbians, provides a cultural affirmation that homosexual orientation in and of itself is not pathologic but is simply part of the human condition.

Prenatal factors. There are extensive data from lower mammals and primates, as well as indirect data from human clinical investigations, that show that the presence of differential amounts of the androgens (hormones that foster male development) in the fetus during a short prenatal period can influence the acquisition and expression of sexually dimorphic behaviour. Androgens are secreted by the fetal gonads and are present in widely differing amounts in all male and female fetuses. Males have about seven times the amount of androgens as have females.

Numerous experiments have reported that female animals exposed to excess fetal androgens behave more like males in various tests of sexually dimorphic behaviour, including a mating preference for the same sex. In 1972 John Money and clinical psychologist Anke Ehrhardt reported that human females exposed to excess fetal androgens as the result of a condition called adrenogenital syndrome behaved throughout childhood with a high degree of energy in rough outdoor play. They also preferred boys as playmates, had little interest in stereotypic activities of girls, and did not verbalize interest in marriage or childbirth. One study found that during adolescence these girls maintained their intense interest in athletics but had difficulty finding female peers with similar interests. Their interest in dating began late, and of those who eventually did begin their sex lives in young adulthood, a significant number were bisexually aroused and some had homosexual experiences. These findings support an earlier report that found that 48 percent of women who were exposed to excess prenatal androgens and developed postnatal androgenization (and therefore were partially virilized—*i.e.*, developed some male secondary sex characteristics, such as a deeper voice or facial hair) had experienced bisexual imagery and 18 percent had had homosexual experiences.

One approach to understanding prenatal influences on sexual orientation is to study the release of hormones from the brains of individuals who are homosexual and transsexual (individuals who, in addition to having homosexual interaction, desire to change their gender). Normally, male and female brains differ, because the female produces hormones in a cyclic manner as part of the menstrual cycle. In 1972 German medical researcher Günter Dörner reported a positive or male-like hormone feedback action in some homosexual women. Similar findings were noted by Lloyd Seyler at the University of Connecticut, who found a male-like hormone response in transsexual women.

Some investigators have found elevated or diminished gonadal hormone levels in some homosexual men and women as compared with those of heterosexuals. The presence of variant hormone levels does not, however, mean that raising or lowering hormone levels in homosexuals will change their sexual orientation. Changing the hormone levels would usually have no effect, although if levels were lowered enough, the threshold for sexual arousal would be raised. A likely explanation for the differing hor-

mone levels is that prenatal androgen may have affected the organization of brain pathways, thereby making some individuals less susceptible to postnatal programming toward heterosexuality.

In the late 1980s and throughout the 1990s biomedical research on homosexuality divided its attention among genetics, brain structure, and hormones. Boston University psychiatry researchers Richard Pillard and James Weinrich concluded in 1986 that homosexuality runs in families. In 1991 Pillard and J. Michael Bailey, a psychologist at Northwestern University, published a study of twins that likewise claimed the "substantial inheritability" of male homosexuality. *Science* magazine, in 1993, published a report by researchers at the National Cancer Institute in Bethesda, Maryland, which concluded that genetic markers on the X chromosome influence homosexual orientation in males.

In 1988 and 1990 neurobiologist D.F. Swaab and associates in The Netherlands reported that they had found a link between homosexuality and the structure of the hypothalamus, a region of the lower brain. Through research at the Salk Institute for Biological Studies, neurobiologist Simon LeVay also posited a correspondence between homosexuality and the hypothalamus in 1991, but his research focused on a different nucleus of the hypothalamus. In 1992 a study at the University of California, Los Angeles (UCLA), concluded that homosexuality may be linked to an area of the brain adjacent to the hypothalamus, the anterior commissure.

Research continues in all these areas. Experts consider that none of the studies so far offers conclusive evidence that biological predispositions to homosexuality exist. There is only the possibility that homosexuality, like heterosexuality, is in part predetermined.

Postnatal factors. The reports by various psychotherapists treating disturbed homosexuals produced the commonly held theory that homosexuality was a disease resulting from pathological influences. It followed that treatment of such pathological influences would result in the spontaneous appearance of the nondisease state—namely, heterosexuality.

The disease model was criticized for various reasons. The clinical evidence has been viewed as biased since only troubled homosexuals came to the psychotherapists' offices. Beginning in the 1940s, Evelyn Hooker, a psychologist at UCLA, studied less biased samples of homosexuals. Her groundbreaking research revealed very divergent—and often healthy and productive—lives. In addition to legitimizing homosexuality as a field of study, Hooker's work contributed to the accumulation of cross-cultural, intersexual, and sociologic data on homosexuality. This and other research led some sexologists to conclude that the human child is born bipotential with regard to erotic orientation. Thus, many began to catalog factors contributing to or inhibiting the unfolding of heterosexuality and homosexuality, rather than labeling the latter a pathological disease.

The rejection of the disease model by many sexologists was most dramatically marked by the vote of the American Psychiatric Association (APA) as early as 1973 to remove "ego-syntonic homosexuality" from the categorization of psychiatric illness in its *Diagnostic and Statistical Manual.* According to the APA, ego-syntonic homosexuality is indicated in people who are accepting of and comfortable with their homosexual orientation; ego-dystonic homosexuality is indicated in people who are concerned with and distressed by their homosexuality.

Regardless of the controversies over whether homosexuality is a "curable" condition, there is increased scientific recognition that the same prenatal and postnatal factors that contribute to heterosexuality also contribute to homosexuality; the difference is the relative contribution of each in an individual's life. Thus, aversive conditioning of the homosexual response (behaviour modification that tries to discourage erotic response to homosexual stimuli) has not led to the spontaneous appearance of heterosexuality, as the disease model would suggest. Indeed, such attempts at so-called "reparative therapy" may indicate an antihomosexual bias on the part of the clinician rather than a useful therapeutic goal. In a *Journal of Homosexuality* article entitled "I'm Your Handyman: A History of Reparative Therapies" (1998), psychiatrist Jack Drescher discussed some of the consequences of the removal of homosexuality from the *Diagnostic and Statistical Manual.* In particular, he noted an increase in political activism against homosexuals as well as new emphases on reparative therapies and transformational ministries (religiously oriented programs that encourage heterosexuality). By contrast, the goal of current therapeutic practice is not to change a person's sexuality but to help clients gain an increased consonance between their sexual orientation, their sense of well-being, and their emotional health.

A likely influence in the development of homosexuality is self-labeling. For example, the experience of homosexual fantasy, while noted above as a common occurrence, might prompt some individuals to label themselves as homosexual and subsequently polarize themselves from heterosexuality. Moreover, early models for the development of homosexuality focused on such issues as impotence and on this concept of polarization in which homosexuals became increasingly oriented (or polarized) away from heterosexuality.

A different model was developed by Australian psychological theorist Vivienne Cass in 1979. She charted the development of male and female homosexual identity through six stages: identity confusion, comparison, tolerance, acceptance, pride, and synthesis. Cass's model depicts identity formation as an individual's movement toward a healthy acceptance of his or her own homosexuality rather than as a polarized rejection of heterosexuality.

Theorists writing about sexual orientation sometimes distinguish between two aspects of human response: *physical* preference (relating to the choice of a sexual partner) and *affectional* preference (relating to emotional connection to another). Physical preferences affect behaviour, while affectional preferences have a psychological component. Because these two aspects of desire are theoretically separable and may be experienced independently (so that a given individual might seek out a person of one sex for sexual fulfillment but prefer someone of the other sex for affectional purposes), the very categories "homosexual" and "heterosexual" may become blurred. Moreover, a person may shift back and forth between same-sex partners and opposite-sex partners for different purposes at different periods of life. The issue is further complicated by the celibate, who may have either same-sex or heterosexual affectional preferences but do not act on their sexual desires.

Homosexual lifestyles. In an attempt to provide empirical data, two major investigations of homosexuality in the United States were conducted in the 1960s and '70s, one by Alan P. Bell and Martin S. Weinberg of the Institute for Sex Research, the other by Masters and Johnson.

Bell and Weinberg interviewed a nonrepresentative sample of some 1,500 persons—homosexuals and heterosexuals, males and females—from the San Francisco Bay area. The major conclusion from their work, *Homosexualities: A Study of Diversity Among Men and Women* (1978), is that it is impossible to predict the social and psychological adjustment of homosexuals or heterosexuals by identifying erotic preferences. Both groups show great diversity in a variety of characteristic categories, including lifestyle, occupation, type of romantic attachment, mental health, happiness, and sexual technique. In some of the characteristic categories, the differences among homosexuals are as wide as or wider than those between homosexuals and heterosexuals.

Regarding gay men, the study showed 14 percent to be "close-coupled" in monogamous bonds. The majority were not involved in monogamous relationships. One-fourth were identified as "open-coupled," meaning that they lived with a special partner but were not monogamous. Slightly less than one quarter organized much of their life around sexual activity and various types of promiscuous behaviour. Another group comprising slightly less than one-fourth of the sample population were classified as "asexual," meaning that these men had little sexual activity. Finally, a smaller number of men were identified as not being content with their homosexuality.

In general, the study indicated that homosexual men were more sexually active than homosexual women were, and the women placed less emphasis on sexual contact. Lesbian cruising (searching for a sexual partner in public places) was infrequent, and women tended to be less interested in impersonal sexual encounters than male homosexuals were. Almost 40 percent of females did not cruise or did so infrequently, and fewer than 20 percent of the females engaged in occasional cruising of gay and lesbian bars and private parties. The females had fewer sexual partners in their lives and were more inclined toward stable monogamous relationships.

Males tended to have several same-sex sexual experiences soon after puberty, whereas the females discovered their homosexuality later in life—usually after a romantic relationship with another female. These data are similar to those published by Marcel T. Saghir and Eli Robins in *Male and Female Homosexuality: A Comprehensive Investigation* (1973); their study noted that most lesbians who had had heterosexual relationships during adolescence found them to be generally less psychosexually satisfying than those reported by heterosexual females. Overall, lesbians in this study were more accepting of and felt less guilty about their homosexuality than the gay men participating in the research.

The study by Masters and Johnson led to the publication of *Homosexuality in Perspective* (1979), a richly detailed account of human sexual behaviour that encompasses various sexual orientations. Their research indicated that the sexual response and physiological capacity to respond to sexual stimuli were not different between homosexuals and heterosexuals. (M.F.S./K.D./Ed.)

BIBLIOGRAPHY

General works. ADRIAN FORSYTH, *A Natural History of Sex* (1986, reissued 1993), treats the role of sex in the natural world. The origins of sex and genetic recombination are considered in JOHN MAYNARD SMITH, *The Evolution of Sex* (1978); and LYNN MARGULIS and DORION SAGAN, *Origins of Sex: Three Billion Years of Genetic Recombination* (1986), and *Mystery Dance: On the Evolution of Human Sexuality* (1991); and the benefits derived by all species from genetic recombination are presented in JAMES L. GOULD and CAROL GRANT GOULD, *Sexual Selection* (1989).

Animals and plants. Courtship and mating are addressed in MARGARET BASTOCK, *Courtship: An Ethological Study* (1967); J.H. PRINCE, *The Universal Urge: Courtship and Mating Among Animals* (1972); ROBERT L. SMITH (ed.), *Sperm Competition and the Evolution of Animal Mating Systems* (1984); and MARK JEROME WALTERS, *The Dance of Life* (1988; also published as *Courtship in the Animal Kingdom,* 1989). The active role of the females in mate selection is investigated in PATRICK BATESON (ed.), *Mate Choice* (1983); EVELYN SHAW and JOAN DARLING, *Female Strategies* (1985); BETTYANN KEVLES, *Females of the Species: Sex and Survival in the Animal Kingdom* (1986); and MARY BATTEN, *Sexual Strategies: How Females Choose Their Mates* (1992).

The sexual systems, mate choices, and pollination strategies of plants are detailed in MARY F. WILLSON, *Plant Reproductive Ecology* (1983); BASTIAAN MEEUSE and SEAN MORRIS, *The Sex Life of Flowers* (1984); and JON LOVETT DOUST and LESLEY LOVETT DOUST (eds.), *Plant Reproductive Ecology: Patterns and Strategies* (1988). Further works treating reproduction in animals and plants may be found in the bibliography to the *Macropædia* article REPRODUCTION AND REPRODUCTIVE SYSTEMS.

Human beings. General information may be found in BENJAMIN B. WOLMAN and JOHN MONEY (eds.), *Handbook of Human Sexuality* (1980, reissued 1993); JAMES LESLIE MCCARY and STEPHEN P. MCCARY, *McCary's Human Sexuality,* 4th ed. (1982); ZIRA DEFRIES, RICHARD C. FRIEDMAN, and RUTH CORN (eds.), *Sexuality: New Perspectives* (1985), a compilation of recent interdisciplinary research; HERANT A. KATCHADOURIAN, *Fundamentals of Human Sexuality,* 5th ed. (1989); JUNE M. REINISCH and RUTH BEASLEY, *The Kinsey Institute New Report on Sex: What You Must Know to Be Sexually Literate* (1990), a summary of current thinking in sex research; WILLIAM H. MASTERS, VIRGINIA E. JOHNSON, and ROBERT C. KOLODNY, *Human Sexuality,* 4th ed. (1992); and JANET SHIBLEY HYDE, *Understanding Human Sexuality,* 5th ed. (1994). A popular treatment of various aspects of sex is found in STEFAN BECHTEL *et al., The Practical Encyclopedia of Sex and Health* (also published as *The Sex Encyclopedia,* 1993). Reference works on human sexuality include ROBERT T. FRANCOEUR, TIMOTHY PERPER, and NORMAN A. SCHERZER (eds.), *A Descriptive Dictionary and Atlas of Sexology* (1991); and MICHAEL A. CARRERA, *The Language of Sex: An A to Z Guide* (1992).

The history of the study of human sexuality is chronicled in PAUL ROBINSON, *The Modernization of Sex: Havelock Ellis, Alfred Kinsey, William Masters, and Virginia Johnson* (1976); VERN L. BULLOUGH, *Sexual Variance in Society and History* (1976); JEFFREY WEEKS, *Sexuality and Its Discontents: Meanings, Myths & Modern Sexualities* (1985), and *Sex, Politics, and Society: The Regulation of Sexuality Since 1800,* 2nd ed. (1989), the latter focusing on Great Britain; PAT CAPLAN (ed.), *The Cultural Construction of Sexuality* (1987); JOHN D'EMILIO and ESTELLE B. FREEDMAN, *Intimate Matters: A History of Sexuality in America* (1988); SANDER L. GILMAN, *Sexuality: An Illustrated History: Representing the Sexual in Medicine and Culture from the Middle Ages to the Age of AIDS* (1989); and JANICE M. IRVINE, *Disorders of Desire: Sex and Gender in Modern American Sexology* (1990).

Significant studies of specifically female or male sexuality in the United States are ALFRED C. KINSEY, WARDELL B. POMEROY, and CLYDE E. MARTIN, *Sexual Behavior in the Human Male* (1948); ALFRED C. KINSEY *et al., Sexual Behavior in the Human Female* (1953, reissued 1973); and SHERE HITE, *The Hite Report: A Nationwide Study on Female Sexuality,* new rev. ed. (1981), and *The Hite Report on Male Sexuality* (1981), although criticism has been leveled at the studies' methodology.

Sexual physiology is comprehensively treated in WILLIAM H. MASTERS and VIRGINIA E. JOHNSON, *Human Sexual Response* (1966, reissued 1986); LORETTA P. HIGGINS and JOELLEN W. HAWKINS, *Human Sexuality Across the Life Span* (1984), a text for nursing practice; and SIMON LEVAY, *The Sexual Brain* (1993), which includes a discussion of the author's report of the size variation of a part of the hypothalamus with respect to sexual orientation among men. WILLIAM H. MASTERS and VIRGINIA E. JOHNSON, *Human Sexual Inadequacy* (1970, reissued 1980), reports behaviour therapy treatments of sexual dysfunction. RICHARD GREEN and JOHN MONEY (eds.), *Transsexualism and Sex Reassignment* (1969), discusses sex change. JOHN MONEY, *Gay, Straight, and In-Between: The Sexology of Erotic Orientation* (1988), analyzes determining physiological, cultural, and personal history factors.

Diseases transmitted through sexual contact are described in CHARLES E. RINEAR, *The Sexually Transmitted Diseases* (1986); and L.C. PARISH and FRIEDRICH GSCHNAIT (eds.), *Sexually Transmitted Diseases: A Guide for Clinicians* (1989).

Treatments of homosexuality in a historical context include VERN L. BULLOUGH, *Homosexuality: A History* (1979); JOHN BOSWELL, *Christianity, Social Tolerance, and Homosexuality: Gay People in Western Europe from the Beginning of the Christian Era to the Fourteenth Century* (1980); and SALVATORE J. LICATA and ROBERT P. PETERSEN (compilers and eds.), *Historical Perspectives on Homosexuality* (1981).

Studies of homosexuality include MARCEL T. SAGHIR and ELI ROBINS, *Male and Female Homosexuality: A Comprehensive Investigation* (1973); ALAN P. BELL and MARTIN S. WEINBERG, *Homosexualities: A Study of Diversity Among Men and Women* (1978); WILLIAM H. MASTERS and VIRGINIA E. JOHNSON, *Homosexuality in Perspective* (1979, reissued 1982); MILTON DIAMOND and ARNO KARLEN, *Sexual Decisions* (1980); and LINDA D. GARNETS (ed.), *Psychological Perspectives On Lesbian and Gay Male Experiences* (1993).

Homosexuality and the humanities are discussed in many texts, including DAVID BELL and GILL VALENTINE (eds.), *Mapping Desire: Geographies of Sexualities* (1995); TERRY CASTLE, *The Apparitional Lesbian: Female Homosexuality and Modern Culture* (1993); EVE KOSOFSKY SEDGWICK, *Between Men: English Literature and Male Homosocial Desire* (1985, reissued 1992), and *Epistemology of the Closet* (1990, reissued 1994); MAJORIE GARBER, *Vice Versa: Bisexuality and the Eroticism of Everyday Life* (1995, reissued 2000); and LILLIAN FADERMAN, *Odd Girls and Twilight Lovers: A History of Lesbian Life in Twentieth-Century America* (1991) and *Surpassing the Love of Men: Romantic Friendship and Love Between Women from the Renaissance to the Present* (1981, reissued 1998).

Also of interest are such scholarly studies as GEORGE WEINBERG, *Society and the Healthy Homosexual* (1972, reissued 1991); RONALD BAYER, *Homosexuality and American Psychiatry: The Politics of Diagnosis* (1981, reprinted 1987), a detailed analysis of the controversy that led to the American Psychiatric Association's decision to remove homosexuality from its list of mental illnesses; JUDD MARMOR (ed.), *Homosexual Behavior: A Modern Reappraisal* (1980), written after the APA's decision and detailing the biological, social science, and clinical views; MICHAEL RUSE, *Homosexuality: A Philosophical Inquiry* (1988, reissued 1990); KENNETH LEWES, *The Psychoanalytic Theory of Male Homosexuality* (1988), an overview from Freud to the present; and RICHARD C. FRIEDMAN, *Male Homosexuality: A Contemporary Psychoanalytic Perspective* (1988).

Shakespeare

Widely regarded as the greatest writer of all time, William Shakespeare occupies a position unique in world literature. Other poets, such as Homer and Dante, and novelists, such as Leo Tolstoy and Charles Dickens, have transcended national barriers; but no writer's living reputation can compare to that of Shakespeare. His plays, written in the late 16th and early 17th centuries for a small repertory theatre, were at the turn of the 21st century performed and read more often and in more countries than ever before. The prophecy of his great contemporary, the poet and dramatist Ben Jonson, that Shakespeare "was not of an age, but for all time" has been fulfilled. He was a writer of great intellectual rapidity, perceptiveness, and poetic power. Other writers have applied their keenness of mind to human beings and their emotions and conflicts, but Shakespeare was astonishingly clever with words and images, so that his mental energy, when applied to intelligible human situations, finds full and memorable expression. As if this were not enough, his chosen art was not remote and bookish but involved vivid stage impersonation, commanding sympathy and inviting vicarious participation.

This article is divided into the following sections:

Shakespeare the man 253
- Life 253
- Early posthumous documentation 254

Shakespeare the poet and dramatist 254
- The intellectual background 254
- Poetic conventions and dramatic traditions 255
- Theatrical conditions 255
- Chronology of Shakespeare's plays 255
- Publication 256

Shakespeare's plays and poems 256
- The early plays 256
- The poems 258
- Plays of the middle and late years 259

Shakespeare's reading 264

Understanding Shakespeare 264
- Questions of authorship 264
- Linguistic, historical, textual, and editorial problems 265
- Literary criticism 265

Shakespeare on film 267

A selected filmography of Shakespeare's works 269

Bibliography 271

Shakespeare the man

LIFE

The amount of factual knowledge available about Shakespeare is surprisingly large for one of his station in life. It is mostly gleaned from documents of an official character: dates of baptisms, marriages, deaths, and burials; wills, conveyances, legal processes, and payments by the court. There are also contemporary allusions to him as a writer that add flesh and blood to the biographical skeleton.

Early life in Stratford. The parish register of Holy Trinity Church, Stratford-upon-Avon, Warwickshire, England, shows that Shakespeare was baptized there on April 26, 1564, but his birthday is traditionally celebrated on April 23. His father, John Shakespeare, was a burgess of the borough, who in 1565 was chosen an alderman and in 1568 bailiff (a position corresponding to mayor). Engaged in various kinds of trade, he appears to have suffered some fluctuations in prosperity. His wife, Mary Arden, of Wilmcote, Warwickshire, came from an ancient family and had inherited some land. (Given the somewhat rigid social distinctions of the 16th century, this marriage must have been a step up the social scale for John Shakespeare.)

Stratford had a grammar school of good quality, and the education there was free, the schoolmaster's salary being paid by the borough. No lists of its 16th-century pupils have survived, but it would be absurd to suppose the town's bailiff did not send his son there. The boy's education would consist mostly of Latin studies—learning to read, write, and speak the language and studying some of the classical historians, moralists, and poets. Shakespeare did not go on to the university, and indeed it is unlikely that the tedious round of university studies, such as logic and rhetoric, would have interested him.

Marriage to Anne Hathaway

Instead, at age 18 he married. The episcopal registry at Worcester preserves a bond dated November 28, 1582, and executed by two yeomen of Stratford, named Sandells and Richardson, as a security to the bishop for the issue of a license for the marriage of William Shakespeare and "Anne Hathaway of Stratford," upon the consent of her friends and upon once asking of the banns. (Anne died in 1623, seven years after Shakespeare. There is good evidence to associate her with a family of Hathaways who inhabited a beautiful farmhouse, now much visited, two miles [3.2 kilometres] from Stratford.) The next date of interest is found in the records of the Stratford church, where a daughter, named Susanna, born to William Shakespeare, was baptized on May 26, 1583. On February 2, 1585, twins were baptized, Hamnet and Judith. (Hamnet, Shakespeare's only son, died 11 years later.)

How Shakespeare spent the next eight years or so, until his name begins to appear in London theatre records, is not known, though there are various conjectures based on the "internal evidence" of his plays.

Career in the theatre. The first reference to Shakespeare in the literary world of London comes in 1592, when a fellow dramatist, Robert Greene, declared in a pamphlet written on his deathbed:

> There is an upstart crow, beautified with our feathers, that with his *Tygers heart wrapt in a Players hide* supposes he is as well able to bombast out a blank verse as the best of you; and, being an absolute *Johannes Factotum,* is in his own conceit the only Shake-scene in a country.

It is difficult to be certain what these words mean, but it is clear that Shakespeare is the object of the sarcasms. When

By courtesy of the Folger Shakespeare Library, Washington, D.C

Shakespeare, first proof of an engraved portrait by Martin Droeshout, from the frontispiece of the First Folio edition of Shakespeare's plays, 1623. In the Folger Shakespeare Library, Washington, D.C.

the book in which they appear (*Greenes, groats-worth of witte, bought with a million of Repentance,* 1592) was published after Greene's death, a mutual acquaintance wrote a preface offering an apology to Shakespeare and testifying to his worth. This preface indicates that Shakespeare was by then making important friends. For, although the puritanical city of London was generally hostile to the theatre, many of the nobility were good patrons of the drama and friends of the actors. Shakespeare seems to have attracted the attention of the young Henry Wriothesley, the 3rd earl of Southampton, to whom he dedicated his first published poems, *Venus and Adonis* and *The Rape of Lucrece.*

Shakespeare seems to have prospered early and tried to retrieve the family's fortunes, at least in part by establishing its gentility. A coat of arms was granted to John Shakespeare in 1596. It can scarcely be doubted that it was William who took the initiative and paid the fees. The coat of arms appears on Shakespeare's monument (constructed before 1623) in the Stratford church. Equally interesting as evidence of Shakespeare's worldly success was his purchase in 1597 of New Place, a large house in Stratford, which as a boy he must have passed every day in walking to school.

The Lord Chamberlain's Men

It is not clear how his career in the theatre began. From about 1594 onward he was an important member of the Lord Chamberlain's Men (called the King's Men after the accession of James I in 1603). They had the best actor, Richard Burbage; they had the best theatre, the Globe; they had the best dramatist, Shakespeare. It is no wonder that the company prospered. Shakespeare became a full-time professional man of his own theatre. For 20 years thereafter he devoted himself assiduously to his art, writing more than a million words of poetic drama of the highest quality.

Private life. Shakespeare had little contact with officialdom, apart from walking—dressed in the royal livery as a member of the King's Men—at the coronation of King James I in 1604. Continuing to look after his financial interests, he bought properties in London and in Stratford. In 1605 he purchased a share (about one-fifth) of the Stratford tithes—a fact that explains why he was eventually buried in the chancel of its parish church. For some time he lodged with a French Huguenot family called Mountjoy, who lived near St. Olave's Church, Cripplegate, London. The records of a lawsuit in May 1612, resulting from a Mountjoy family quarrel, show Shakespeare as giving evidence in a genial way and as interesting himself generally in the family's affairs.

No letters written by Shakespeare have survived, but a private letter to him happened to get caught up with some official transactions of the town of Stratford and so has been preserved in the borough archives. It was written by one Richard Quiney and addressed by him from the Bell Inn in Carter Lane, London. On one side of the paper is inscribed: "To my loving good friend and countryman, Mr. Wm. Shakespeare, deliver these." Apparently Quiney thought his fellow Stratfordian a person to whom he could apply for the loan of £30—a large sum in Elizabethan money. Nothing further is known about the transaction, though it is of some interest that 18 years later Quiney's son Thomas became the husband of Judith, Shakespeare's second daughter.

Shakespeare's will, which was made on March 25, 1616, is a long and detailed document. It entailed his quite ample property on the male heirs of his elder daughter, Susanna. (Both his daughters were then married, one to the aforementioned Thomas Quiney and the other to John Hall, a respected physician of Stratford.) As an afterthought, he bequeathed his "second-best bed" to his wife; no one can be certain what this notorious legacy means. The testator's signatures to the will are apparently in a shaky hand. Perhaps Shakespeare was already ill. He died on April 23, 1616. No name was inscribed on his gravestone in the chancel of the parish church of Stratford-upon-Avon. Instead, these lines, possibly his own, appeared:

> Good friend, for Jesus' sake forbear
> To dig the dust enclosed here.
> Blest be the man that spares these stones,
> And curst be he that moves my bones.

EARLY POSTHUMOUS DOCUMENTATION

Within a few years, a monument was erected on the chancel wall. It seems to have existed by 1623. Its epitaph, written in Latin, attributes to Shakespeare the worldly wisdom of Nestor, the genius of Socrates, and the poetic art of Virgil. This apparently was how his contemporaries in Stratford-upon-Avon wished him to be remembered.

The tributes of his colleagues. Shakespeare's plays remained a major part of the repertory of the King's Men until the closing of the theatres in 1642. To William Drummond of Hawthornden in 1619 Ben Jonson said that Shakespeare "wanted art." But, when Jonson came to write his splendid poem prefixed to the Folio edition of Shakespeare's plays in 1623, he rose to the occasion with stirring words of praise:

> Triumph, my Britain, thou hast one to show
> To whom all scenes of Europe homage owe.
> He was not of an age, but for all time!

Besides almost retracting his earlier gibe about Shakespeare's lack of art, he gives testimony that Shakespeare's personality was to be felt, by those who knew him, in his poetry—that the style was the man. Jonson also reminded his readers of the strong impression the plays had made upon Queen Elizabeth I and King James I at court performances:

> Sweet Swan of Avon, what a sight it were
> To see thee in our waters yet appear,
> And make those flights upon the banks of Thames
> That so did take Eliza and our James!

Shakespeare seems to have been on affectionate terms with his theatre colleagues. His fellow actors John Heminge and Henry Condell (who, with Burbage, were remembered in his will) dedicated the First Folio of 1623 to the earl of Pembroke and the earl of Montgomery, explaining that they had collected the plays "without ambition either of self-profit or fame; only to keep the memory of so worthy a friend and fellow alive as was our Shakespeare."

Anecdotes and documents. Seventeenth-century antiquaries began to collect anecdotes about Shakespeare, but no serious life was written until 1709, when Nicholas Rowe tried to assemble information from all available sources. There were local traditions at Stratford: witticisms and lampoons of local characters; scandalous stories of drunkenness and sexual escapades. About 1661 the vicar of Stratford wrote in his diary: "Shakespeare, Drayton, and Ben Jonson had a merry meeting, and it seems drank too hard; for Shakespeare died of a fever there contracted." On the other hand, the antiquary John Aubrey wrote in some notes about Shakespeare: "He was not a company keeper; lived in Shoreditch; wouldn't be debauched, and, if invited to, writ he was in pain." Richard Davies, archdeacon of Lichfield, reported, "He died a papist." How much trust can be put in such a story is uncertain. In the early 18th century, a story appeared that Queen Elizabeth had obliged Shakespeare "to write a play of Sir John Falstaff in love" and that he had performed the task (*The Merry Wives of Windsor*) in a fortnight. There are other stories, all of uncertain authenticity and some mere fabrications.

Anecdotes about Shakespeare

When serious scholarship began in the 18th century, documents were discovered. Shakespeare's will was found in 1747 and his marriage license in 1836. The documents relating to the Mountjoy lawsuit already mentioned were found and printed in 1910. It is conceivable that further documents of a legal nature may yet be discovered, but modern scholarship is more concerned to study Shakespeare in relation to his social environment, both in Stratford and in London. This is not easy, because the author and actor lived a somewhat detached life: a respected tithe-owning country gentleman in Stratford, perhaps, but a rather rootless artist in London. (J.R.Br./T.Sp./Ed.)

Shakespeare the poet and dramatist

THE INTELLECTUAL BACKGROUND

Shakespeare lived at a time when ideas and social structures established in the Middle Ages still informed human thought and behaviour. Queen Elizabeth I was God's

deputy on earth, and lords and commoners had their due places in society under her, with responsibilities up through her to God and down to those of more humble rank. The order of things, however, did not go unquestioned. Atheism was still considered a challenge to the beliefs and way of life of a majority of Elizabethans, but the Christian faith was no longer single. Rome's authority had been challenged by Martin Luther, John Calvin, a multitude of small religious sects, and, indeed, the English church itself. Royal prerogative was challenged in Parliament; the economic and social orders were disturbed by the rise of capitalism, by the redistribution of monastic lands under Henry VIII, by the expansion of education, and by the influx of new wealth from discovery of new lands.

Ideas of the time

An interplay of new and old ideas was typical of the time: official homilies exhorted the people to obedience; the Italian political theorist Niccolò Machiavelli was expounding a new, practical code of politics that caused Englishmen to fear the Italian "Machiavillain" and yet prompted them to ask what men do, rather than what they should do. In *Hamlet,* disquisitions—on man, belief, a "rotten" state, and times "out of joint"—clearly reflect a growing disquiet and skepticism. The translation of Montaigne's *Essays* in 1603 gave further currency, range, and finesse to such thought, and Shakespeare made direct and significant quotations from the essays in *The Tempest.* In philosophical inquiry the question "how?" became the impulse for advance, rather than the traditional "why?" of Aristotle. Shakespeare's plays written between 1603 and 1606 unmistakably reflect a new, Jacobean distrust. James I, who, like Elizabeth, claimed divine authority, was far less able than she to maintain the authority of the throne. The so-called Gunpowder Plot (1605) showed a determined challenge by a small minority in the state; James's struggles with the House of Commons in successive Parliaments, in addition to indicating the strength of the "new men," also revealed the inadequacies of the administration.

POETIC CONVENTIONS AND DRAMATIC TRADITIONS

The Latin comedies of Plautus and Terence were familiar in Elizabethan schools and universities, and English translations or adaptations of them were occasionally performed by students. Seneca's rhetorical and sensational tragedies, too, had been translated and often imitated. But there was also a strong native dramatic tradition deriving from the medieval miracle plays, which had continued to be performed in various towns until forbidden during Elizabeth's reign. This native drama had been able to assimilate French popular farce, clerically inspired morality plays on abstract themes, and interludes or short entertainments that made use of the "turns" of individual clowns and actors. Although Shakespeare's immediate predecessors were known as "University wits," their plays were seldom structured in the manner of those they had studied at Oxford or Cambridge; instead, they used and developed the more popular narrative forms.

Changes in the English language

Changes in language. The English language at this time was changing and extending its range. The poet Edmund Spenser led with the restoration of old words, and schoolmasters, poets, sophisticated courtiers, and travelers all brought further contributions from France, Italy, and the Roman classics, as well as from farther afield. Helped by the growing availability of cheaper, printed books, the language began to become standardized in grammar and vocabulary and, more slowly, in spelling. Ambitious for a European and permanent reputation, the essayist and philosopher Francis Bacon wrote in Latin as well as in English; but, if he had lived only a few decades later, even he might have had total confidence in his own tongue.

Shakespeare's literary debts. In Shakespeare's earlier works his debts stand out clearly: to Plautus for the structure of *The Comedy of Errors;* to the poet Ovid and to Seneca for rhetoric and incident in *Titus Andronicus;* to morality drama for a scene in which a father mourns his dead son, and a son his father, in *Henry VI;* to Marlowe for sentiments and characterization in *Richard III* and *The Merchant of Venice;* to the Italian popular tradition of commedia dell'arte for characterization and dramatic style in *The Taming of the Shrew;* and so on. Soon, however, there was no line between their effects and his. In *The Tempest* (perhaps the most original of all his plays in form, theme, language, and setting) folk influences may also be traced, together with a newer and more obvious debt to a courtly diversion known as the masque, as developed by Ben Jonson and others at the court of King James.

THEATRICAL CONDITIONS

The Globe and its predecessor, the Theatre, were public playhouses run by the Chamberlain's Men, a leading theatre company of which Shakespeare was a member. Almost all classes of citizens, except the Puritans, came to them for afternoon entertainment. The players were also summoned to court, to perform before the monarch and assembled nobility. In the summer they toured the provinces, and on occasion they performed at London's Inns of Court (associations of law students), at universities, and in great houses. Popularity led to an insatiable demand for plays: early in 1613 the King's Men—as the Chamberlain's Men were then known—could present "fourteen several plays." The theatre soon became fashionable, too, and in 1608–09 the King's Men started to perform on a regular basis at the Blackfriars, a "private" indoor theatre where high admission charges assured the company a more select and sophisticated audience for their performances.

Shakespeare's first associations with the Chamberlain's Men seem to have been as an actor. He is not known to have acted after 1603, and tradition gives him only secondary roles, such as the ghost in *Hamlet* and Adam in *As You Like It,* but his continuous association must have given him direct working knowledge of all aspects of theatre. Numerous passages in his plays show conscious concern for theatre arts and audience reactions. Prospero in *The Tempest* speaks of the whole of life as a kind of "revels," or theatrical show, that, like a dream, will soon be over. The Duke of York in *Richard II* is conscious of how

Shakespeare's knowledge of the stage

> . . . in a theatre, the eyes of men,
> After a well-graced actor leaves the stage
> Are idly bent on him that enters next,
> Thinking his prattle to be tedious.

And Hamlet gives expert advice to visiting actors in the art of playing.

In Shakespeare's day, there was little time for group rehearsals, and actors were given the words of only their own parts. The crucial scenes in Shakespeare's plays, therefore, are between two or three characters only or else are played with one character dominating a crowded stage. Female parts were written for young male actors or boys, so Shakespeare did not often write big roles for them or keep them actively engaged onstage for lengthy periods. Writing for the clowns of the company—who were important popular attractions in any play—presented the problem of allowing them to use their comic personalities and tricks and yet have them serve the immediate interests of theme and action.

CHRONOLOGY OF SHAKESPEARE'S PLAYS

Despite much scholarly argument, it is often impossible to date a given play precisely. But there is a general consensus, especially for plays written 1585–1601, 1605–07, and 1609 onward. The following list of first performances is based on external and internal evidence, on general stylistic and thematic considerations, and on the observation that an output of no more than two plays a year seems to have been established in those periods when dating is rather clearer than others.

1589–94	*1 Henry VI, 2 Henry VI, 3 Henry VI, Richard III, The Comedy of Errors, The Two Gentlemen of Verona, Love's Labour's Lost, The Taming of the Shrew, Titus Andronicus*
1594–96	*King John, Romeo and Juliet*
1595–96	*A Midsummer Night's Dream, Richard II*
1596–97	*The Merchant of Venice, 1 Henry IV*
1597–98	*2 Henry IV*
1597–1601	*The Merry Wives of Windsor*
1598–99	*Much Ado About Nothing*

1598–1600	*As You Like It*
1599	*Henry V, Julius Caesar*
1599–1601	*Hamlet*
1600–02	*Twelfth Night*
1601–02	*Troilus and Cressida*
1601–05	*All's Well That Ends Well*
1603–04	*Measure for Measure, Othello*
1605–06	*King Lear*
1605–08	*Timon of Athens*
1606–07	*Macbeth, Antony and Cleopatra*
1606–08	*Pericles*
1608	*Coriolanus*
1608–10	*Cymbeline*
1609–11	*The Winter's Tale*
1611	*The Tempest*
1613	*Henry VIII*
1613–14	*The Two Noble Kinsmen*

Shakespeare's two narrative poems, *Venus and Adonis* and *The Rape of Lucrece,* can be dated with certainty to the years when the plague stopped dramatic performances in London, in 1592 and 1593–94, respectively, just before their publication. But the sonnets offer many and various problems; they cannot have been written all at one time, and most scholars set them within the period 1593–1600. "The Phoenix and the Turtle" can be dated 1600–01.

PUBLICATION

Acting companies in London during the Renaissance were perennially in search of new plays. Ordinarily they paid individual playwrights for them, on a freelance, piecework basis. Shakespeare was an important exception; as a member of the Lord Chamberlain's and then the King's men, he wrote for his company as a sharer in their capitalist enterprise.

The companies were not eager to sell their plays to publishers. At certain times, however, they might have been impelled to do so: when a company disbanded or was put into enforced inactivity by visitations of the plague, or when the plays were no longer current. (The companies owned the plays; the individual authors had no intellectual property rights in them once the plays had been sold to the actors.)

Such plays were usually published in quarto form—that is, printed on both sides of large sheets of paper with four printed pages on each side. When the sheet was folded twice and bound, it yielded eight printed pages to each "gathering." A few plays were printed in octavo, with the sheet folded thrice and yielding 16 smaller printed pages to each gathering.

Differing versions of the plays

Half of Shakespeare's plays were printed in quarto (at least one in octavo) during his lifetime. Occasionally, a play was issued in a seemingly unauthorized volume—that is, not having been regularly sold by the company to the publisher. The acting company might then have commissioned its own authorized version. The quarto title page of *Romeo and Juliet* (1599), known today as the second quarto, declares on its title page that it is "Newly corrected, augmented, and amended, as it hath been sundry times publicly acted by the Right Honorable the Lord Chamberlain His Servants." The second quarto of *Hamlet* (1604–05) similarly advertises itself as "Newly imprinted and enlarged to almost as much again as it was, according to the true and perfect copy." Indeed, the first quarto of *Hamlet* (1603) is considerably shorter than the second, and the first quarto of *Romeo and Juliet* lacks some 800 lines found in its successor. Both contain what appear to be misprints or other errors that are then corrected in the second quarto.

The status of these and other seemingly unauthorized editions is much debated today. The older view of A.W. Pollard, W.W. Greg, Fredson Bowers, and other practitioners of the so-called New Bibliography generally regards these texts as suspect and perhaps pirated, either by unscrupulous visitors to the theatre or by minor actors who took part in performance and who then were paid to reconstruct the plays from memory. The unauthorized texts do contain elements that sound like the work of eyewitnesses or actors (and are valuable for that reason). In some instances, the unauthorized text is notably closer to the authorized text when certain minor actors are onstage than at other times, which suggests that these actors may have been involved in a memorial reconstruction. The plays *2 Henry VI* and *3 Henry VI* originally appeared in shorter versions that may have been memorially reconstructed by actors.

A revisionary school of textual criticism that gained favour in the latter part of the 20th century argued that these texts might have been earlier versions with their own theatrical rationale and that they should be regarded as part of a theatrical process by which the plays evolved onstage. Certainly the situation varies from quarto to quarto, and unquestionably the unauthorized quartos are valuable to the understanding of stage history.

The first collected edition

Several years after Shakespeare died in 1616, colleagues of his in the King's Men, John Heminge and Henry Condell, undertook the assembling of a collected edition. It appeared in 1623 as *Mr. William Shakespeare's Comedies, Histories, and Tragedies, published according to the true original copies.* It did not contain the poems and left out *Pericles* as perhaps of uncertain authorship; nor did it include *The Two Noble Kinsmen, Edward III,* or the portion of *The Book of Sir Thomas More* that Shakespeare may have contributed. It did nonetheless include 36 plays, half of them appearing in print for the first time.

Heminge and Condell had the burdensome task of choosing what materials to present to the printer, for they had on hand a number of authorial manuscripts, other documents that had served as promptbooks for performance (these were especially valuable, since they bore the license for performance), and some 18 plays that had appeared in print. Fourteen of these had been published in what the editors regarded as more or less reliable texts, though only two were used unaltered. Much was discovered by textual scholarship after Heminge and Condell did their original work, and the result was a considerable revision in what came to be regarded as the best choice of original text from which an editor ought to work. In plays published in both folio and quarto (or octavo) format, the task of choosing was immensely complicated. *King Lear* especially became a critical battleground in which editors argued for the superiority of various features of the 1608 quarto or the folio text. The two differ substantially and must indeed represent different stages of composition and of staging, so that both are germane to an understanding of the play's textual and theatrical history. The same is true of *Hamlet,* with its unauthorized quarto of 1603, its corrected quarto of 1604–05, and the folio text, all significantly at variance with one another. There are several other plays in which the textual relationship of quarto to folio is highly problematic. Information on these is readily available in critical editions of Shakespeare's plays and poems.

(J.R.Br./T.Sp./D.Bev.)

Shakespeare's plays and poems

THE EARLY PLAYS

Shakespeare arrived in London probably sometime in the late 1580s. He was in his mid-20s. It is not known how he got started in the theatre, or for what acting companies he wrote his early plays. Indicating a time of apprenticeship, these plays show a more direct debt to London dramatists of the 1580s and to classical examples than do his later works.

Titus Andronicus. *Titus Andronicus* (*c.* 1589–92) is a case in point. As Shakespeare's first full-length tragedy, it owes much of its theme, structure, and language to Thomas Kyd's huge success in the late 1580s with *The Spanish Tragedy.* Kyd had hit on the formula of adopting the dramaturgy of Seneca (the younger), the great Stoic philosopher and statesman, to the needs of a burgeoning new London theatre. The result was the revenge tragedy, an astonishingly successful genre that was to be refigured in *Hamlet* and many other revenge plays. Shakespeare also borrowed a leaf from his great contemporary Christopher Marlowe. The Vice-like protagonist of Marlowe's *The Jew of Malta,* Barabas, may have inspired Shakespeare in his depiction of the villainous Aaron the Moor in *Titus Andronicus.*

The Senecan model offered Kyd, and then Shakespeare, a story of bloody revenge, occasioned originally by the murder or rape of a person whose near relatives (fathers, sons, brothers) are bound by sacred oath to revenge the atrocity. The avenger must proceed with caution, since his opponent is canny, secretive, and ruthless. The avenger becomes mad or feigns madness to cover his intent. He becomes more and more ruthless himself as he moves toward his goal of vengeance. At the same time, he is hesitant, being deeply distressed by ethical considerations. An ethos of revenge is opposed to one of Christian forbearance. The avenger may see the spirit of the person whose wrongful death he must avenge. He employs the device of a play within a play in order to accomplish his aims. The play ends in a bloodbath and a vindication of the avenger. Evident in this model is the story of Titus Andronicus, whose sons are butchered and whose daughter is ravished, as well as the story of Hamlet and still others.

The early romantic comedies. Other than *Titus Andronicus,* Shakespeare did not experiment with formal tragedy in his early years. (Though his English history plays from this period portrayed tragic events, their theme was focused elsewhere.) The young playwright was drawn more quickly into comedy, and with more immediate success. For this his models include the dramatists Robert Greene and John Lyly, along with Thomas Nashe. The result is a genre recognizably and distinctively Shakespearean, even if he learned a lot from Greene and Lyly: the romantic comedy. As in the work of his models, Shakespeare's early comedies revel in stories of amorous courtship in which a plucky and admirable young woman (played by a boy actor) is paired off against her male wooer. Julia, one of two young heroines in *The Two Gentlemen of Verona* (*c.* 1590–94), disguises herself as a man in order to follow her lover, Proteus, when he is sent away from Verona to Milan. Proteus (appropriately named), she discovers, is paying far too much attention to Sylvia, the beloved of Proteus's best friend, Valentine. Love and friendship thus do battle for the divided loyalties of the erring male until the generosity of his friend and, most of all, the enduring chaste loyalty of the two women bring Proteus to his senses. The motif of the young woman disguised as a male was to prove invaluable to Shakespeare in subsequent romantic comedies, including *The Merchant of Venice, As You Like It,* and *Twelfth Night.* As is generally true of Shakespeare, he derived the essentials of his plot from a narrative source, in this case a long Spanish prose romance, the *Diana* of Jorge de Montemayor.

Shakespeare's most classically inspired early comedy is *The Comedy of Errors* (*c.* 1589–94). Here he turned particularly to Plautus's farcical play called the *Menaechmi* (*Twins*). The story of one twin (Antipholus) looking for his lost brother, accompanied by a clever servant (Dromio) whose twin has also disappeared, results in a farce of mistaken identities that also thoughtfully explores issues of identity and self-knowing. The young women of the play, one the wife of Antipholus of Ephesus (Adriana) and the other her sister (Luciana), engage in meaningful dialogue on issues of wifely obedience or autonomy. Marriage resolves these difficulties at the end, as is routinely the case in Shakespearean romantic comedy, but not before the plot complications have tested the characters' needs to know who they are and what men and women ought to expect from one another.

Courtship and marriage in the early romantic comedies

Shakespeare's early romantic comedy most indebted to John Lyly is *Love's Labour's Lost* (*c.* 1588–97), a confection set in the never-never land of Navarre, where the King and his companions are visited by the Princess of France and her ladies-in-waiting on a diplomatic mission that soon devolves into a game of courtship. As is often the case in Shakespearean romantic comedy, the young women are sure of who they are and whom they intend to marry; one cannot be certain that they ever really fall in love, since they begin by knowing what they want. The young men, conversely, fall all over themselves in their comically futile attempts to eschew romantic love in favour of more serious pursuits. Shakespeare brilliantly portrays male discomfiture and female self-assurance as he explores the treacherous but desirable world of sexual attraction, while the verbal gymnastics of the play emphasize the wonder and the delicious foolishness of falling in love.

The Taming of the Shrew

In *The Taming of the Shrew* (*c.* 1590–93), Shakespeare employs a device of multiple plotting that is to become a standard feature of his romantic comedies. In one plot, derived from Ariosto's *I Suppositi* (*Supposes,* as it had been translated into English by George Gascoigne), a young woman (Bianca) carries on a risky courtship with a young man who appears to be a tutor, much to the dismay of her father, Baptista Minola, who hopes to marry her to a wealthy suitor of his own choosing. Eventually the mistaken identities are straightened out, establishing the presumed tutor as Lucentio, wealthy and suitable enough. Simultaneously, Bianca's shrewish sister, Kate, denounces (and terrorizes) all men. Bianca's suitors commission the self-assured Petruchio to pursue Kate so that Bianca, the younger sister, will then be free to wed. The wife-taming plot is itself based on folktale and ballad tradition in which men ensure their ascendancy in the marriage relationship by beating their wives into submission. Shakespeare transforms this raw, antifeminist material into a study of the struggle for dominance in the marriage relationship. And, whereas he does opt in this play for male triumph over the female, he gives to Kate a sense of humour that enables her to see how she is to play the game to her own advantage as well. She is, arguably, happy at the end with a relationship based on wit and companionship, whereas Bianca turns out to be simply spoiled.

Shakespeare "invents" the English history play

The early histories. In Shakespeare's explorations of English history, as in romantic comedy, he put his distinctive mark on a genre and made it his. The genre was, moreover, an unusual one. There was no definition of an English history play, and there were no aesthetic rules regarding its shaping. The ancient classical world had recognized two broad categories of genre, comedy and tragedy. (This account leaves out more specialized genres such as the satyr play.) Aristotle and other critics, including Horace, had evolved, over centuries, classical definitions. Tragedy dealt with the disaster-struck lives of great persons, was written in elevated verse, and took as its setting a mythological and ancient world of gods and heroes: Agamemnon, Theseus, Oedipus, Medea, and the rest. Pity and terror were the prevailing emotional responses in plays that sought to understand, however imperfectly, the will of the supreme gods. Classical comedy, conversely, dramatized the everyday. Its chief figures were citizens of Athens and Rome—householders, courtesans, slaves, scoundrels, and so forth. The humour was immediate, contemporary, topical; the lampooning was satiric, even savage. Members of the audience were invited to look at mimetic representations of their own lives and laugh at greed and folly.

The English history play had no such ideal theoretical structure. It was an existential invention: the dramatic treatment of recent English history. It might be tragic or comic or, more commonly, a hybrid. Polonius's list of generic possibilities captures the ludicrous potential for endless hybridizations: "tragedy, comedy, history, pastoral, pastoral-comical, historical-pastoral, tragical-historical, tragical-comical-historical-pastoral," and so on (*Hamlet,* 2.2.397–99). Shakespeare's history plays were so successful in the 1590s London theatre that the editors of Shakespeare's complete works, in 1623, chose to group his dramatic output under three headings: comedies, histories, and tragedies. The genre established itself by sheer force of its compelling popularity.

Shakespeare, in 1590 or thereabouts, had really only one viable model for the English history play, an anonymous and sprawling drama called *The Famous Victories of Henry the Fifth* (1583–88) that told the saga of Henry IV's son, Prince Hal, from the days of his adolescent rebellion down through his victory over the French at the battle of Agincourt in 1415—in other words, the material that Shakespeare would later use in writing three major plays, *1 Henry IV, 2 Henry IV,* and *Henry V.* Shakespeare chose to start not with Prince Hal but with more recent history in the reign of Henry V's son Henry VI and with the civil wars that saw the overthrow of Henry VI by Edward IV and then the accession to power in 1483 of Richard III.

This material proved to be so rich in themes and dramatic conflicts that he wrote four plays on it, a "tetralogy" extending from *Henry VI* in three parts (*c.* 1589–92) to *Richard III* (*c.* 1592–94).

These plays were immediately successful. Contemporary references indicate that audiences thrilled to the story (in *1 Henry VI*) of the brave Lord Talbot, doing battle in France against the witch Joan of Arc and her lover, the French Dauphin, but being undermined in his heroic effort by effeminacy and corruption at home. Henry VI himself is, as Shakespeare portrays him, a weak king, raised to the kingship by the early death of his father, Henry V, incapable of controlling factionalism in his court and enervated personally by his infatuation with a dangerous Frenchwoman, Margaret of Anjou. Henry VI is cuckolded by his wife and her lover, the Duke of Suffolk, and (in *2 Henry VI*) proves unable to defend his virtuous uncle, the Duke of Gloucester, against opportunistic enemies. The result is civil unrest, lower-class rebellion, and eventually all-out civil war between the Lancastrian faction, nominally headed by Henry VI, and the Yorkist claimants under the leadership of Edward IV and his brothers. *Richard III* completes the saga with its account of the baleful rise of Richard of Gloucester through the murdering of his brother the Duke of Clarence and of Edward IV's two sons, who were also Richard's nephews. Richard's tyrannical reign yields eventually and inevitably to the newest and most successful claimant of the throne, Henry Tudor, earl of Richmond. This is the man who becomes Henry VII, scion of the Tudor dynasty and grandfather of Queen Elizabeth I, who reigned from 1558 to 1603 and hence during the entire first decade and more of Shakespeare's productive career.

Civil unrest in the history plays

The Shakespearean English history play told of the country's history at a time when the English nation was struggling with its own sense of national identity and experiencing a new sense of power. Queen Elizabeth had brought stability and a relative freedom from war to her decades of rule. She had held at bay the Catholic powers of the Continent, notably Philip II of Spain, and, with the help of a storm at sea, had fought off Philip's attempts to invade her kingdom with the great Spanish Armada of 1588. In England the triumph of the nation was viewed universally as a divine deliverance. A new edition of Raphael Holinshed's *Chronicles,* published in 1587, was at hand as a vast source for Shakespeare's historical playwriting. It, too, celebrated the emergence of England as a major Protestant power, led by a popular and astute monarch.

From the perspective of the 1590s, the history of the 15th century also seemed newly pertinent. England had emerged from a terrible civil war in 1485 with Henry Tudor's victory over Richard III at the Battle of Bosworth Field. The chief personages of the English Civil Wars—Henry Tudor, Richard III, the Duke of Buckingham, Hastings, Rivers, Gray, and many more—were very familiar to contemporary English readers.

Fictional aspects of the history plays

Because these historical plays of Shakespeare in the early 1590s were so intent on telling the saga of emergent nationhood, they exhibit a strong tendency to identify villains and heroes. Shakespeare is writing dramas, not schoolbook texts, and he freely alters dates and facts and emphases. Lord Talbot in *1 Henry VI* is a hero because he dies defending English interests against the corrupt French. In *2 Henry VI* Humphrey, duke of Gloucester, is cut down by opportunists because he represents the best interests of the commoners and the nation as a whole. Most of all, Richard of Gloucester is made out to be a villain epitomizing the very worst features of a chaotic century of civil strife. He foments strife, lies, and murders and makes outrageous promises he has no intention of keeping. He is a brilliantly theatrical figure because he is so inventive and clever, but he is also deeply threatening. The real Richard was no such villain, it seems; at least, his politically inspired murders were no worse than the systematic elimination of all opposition by his successor, the historical Henry VII. The difference is that Henry VII lived to commission historians to tell the story his way, whereas Richard lost everything through defeat. As founder of the Tudor dynasty and grandfather of Queen Elizabeth, Henry VII could command a respect that even Shakespeare was bound to honour, and accordingly the Henry Tudor that he portrays at the end of *Richard III* is a God-fearing patriot and loving husband of the Yorkist princess who is to give birth to the next generation of Tudor monarchs.

Richard III is a tremendous play, both in length and in the bravura depiction of its titular protagonist. It is called a tragedy on its original title page, as are other of these early English history plays. Certainly they present brutal deaths and instructive falls of great men from positions of high authority to degradation and misery. Yet these plays are not tragedies in the classical sense of the term. They contain so much else, and notably they end on a major key: the accession to power of the Tudor dynasty that will give England its great years under Elizabeth. The story line is one of suffering and of eventual salvation, of deliverance by mighty forces of history and of divine oversight that will not allow England to continue to suffer once it has returned to the true path of duty and decency. In this important sense, the early history plays are like tragicomedies, or romances.

THE POEMS

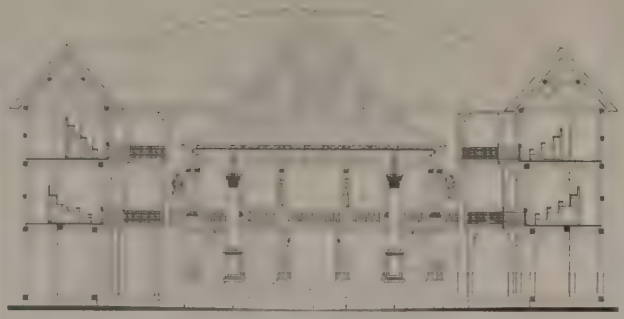
Shakespeare seems to have wanted to be a poet as much as he sought to succeed in the theatre. His plays are wonderfully and poetically written, often in blank verse. And when he experienced a pause in his theatrical career about 1592–94, the plague having closed down much theatrical activity, he wrote poems. *Venus and Adonis* (1593) and *The Rape of Lucrece* (1594) are the only works that Shakespeare seems to have shepherded through the printing process. Both owe a good deal to Ovid, the classical poet whose writings Shakespeare encountered repeatedly in school. These two poems are the only works for which he wrote dedicatory prefaces. Both are to Henry Wriothesley, earl of Southampton. This young man, a favourite at court, seems to have encouraged Shakespeare and to have served for a brief time at least as his sponsor. The dedication to the second poem is measurably warmer than the first. An unreliable tradition supposes that Southampton gave Shakespeare the stake he needed to buy into the newly formed Lord Chamberlain's acting company in 1594. Shakespeare became an actor-sharer, one of the owners in a capitalist enterprise that shared the risks and the gains among them. This company succeeded brilliantly; Shakespeare and his colleagues, including Richard Burbage, John Heminge, Henry Condell, and Will Sly, became wealthy through their dramatic presentations.

Shakespeare may also have written at least some of his sonnets to Southampton, beginning in these same years of 1593–94 and continuing on through the decade and later. The question of autobiographical basis in the sonnets is much debated, but Southampton at least fits the portrait of a young gentleman who is being urged to marry and produce a family. (Southampton's family was eager that he do just this.) Whether the account of a strong loving relationship between the poet and his gentleman friend is autobiographical is more difficult still to determine. As a narrative, the sonnet sequence tells of strong attachment, of jealousy, of grief at separation, of joy at being together and sharing beautiful experiences. The emphasis on the importance of poetry as a way of eternizing human achievement and of creating a lasting memory for the poet himself is appropriate to a friendship between a poet of modest social station and a friend who is better born. When the sonnet sequence introduces the so-called "Dark Lady," the narrative becomes one of painful and destructive jealousy. Scholars do not know the order in which the sonnets were composed—Shakespeare seems to have had no part in publishing them—but no order other than the order of publication has been proposed, and as the sonnets stand they tell a coherent and disturbing tale. The poet experiences sex as something that fills him with revulsion and remorse, at least in the lustful circumstances in which he encounters it. His attachment to the young

Shakespeare's birthplace at Stratford-upon-Avon, Warwickshire, England; watercolour by Phoebe Dighton, 1834

William Shakespeare, detail of an oil painting attributed to John Taylor, *c.* 1610. The portrait is called the "Chandos Shakespeare" because it once belonged to the duke of Chandos.

Cardinal Wolsey in *Henry VIII,* as portrayed by Sir Henry Irving at the Lyceum Theatre, London, 1892



Cross section of the New Globe Theatre, London, built 1987–97, showing the galleries on either side, the stage, and the *frons scenae,* or stage wall.

The shrew Katharina in *The Taming of the Shrew,* as presented by the 15-year-old Laurence Olivier in the Shakespeare Festival Theatre, Stratford-upon-Avon, 1922.

Richard III in *Richard III,* as performed by Sir David Garrick at Drury Lane, London, 1759; oil painting on canvas by Francis Hayman, 1760

Copperplate engraving of the Globe Theatre, Bankside, where Shakespeare's plays were performed after 1599.

Plate 1 (Top left) By
Shakespeare
The National
Hulton Gett
by Jon Gr
left and bottom ng
(bottom centre) National Pl
The Art Archive; photograph

Cleopatra in *Antony and Cleopatra,* as portrayed by Vivien Leigh in the St. James Theatre, London, 1951. At the time of this run, Leigh was married to the play's director, Laurence Olivier.

Desdemona and Othello in *Othello,* as played by Peggy Ashcroft and American actor Paul Robeson in the Savoy Theatre, London, 1930.

Richard III in *Richard III* (set in 1930s London), as performed by Sir Ian McKellen in a 1995 film adaptation of the play.

Beatrice and Benedick, as performed by Emma Thompson and Kenneth Branagh in Branagh's film adaptation of *Much Ado About Nothing,* 1993.

Hamlet in *Hamlet,* as played by John Gielgud in the New Theatre, London, 1934, a program he also directed. He performed the role many times from 1930 to 1946

River Phoenix as Mike Waters and Chira Caselli as Carmella in Gus Van Sant's film *My Own Private Idaho* (1991), adapted from Shakespeare's *Henry IV, Part II.*

Plate 2. (Top left) Mander and Mitchenson Theatre Collection, (centre left) United Artists/The Kobal Collection, (bottom left) Hulton Getty/Archive Photos, (top right) Bettmann/Corbis, (centre right) Hulton Archive, (bottom right) Everett Collection

man is a love relationship that sustains him at times more than the love of the Dark Lady can, and yet this loving friendship also dooms the poet to disappointment and self-hatred. Whether the sequence reflects any circumstances in Shakespeare's personal life, it certainly is told with an immediacy and dramatic power that bespeak an extraordinary gift for seeing into the human heart and its sorrows.

PLAYS OF THE MIDDLE AND LATE YEARS

Romantic comedies. In the second half of the 1590s, Shakespeare brought to perfection the genre of romantic comedy that he had helped to invent. *A Midsummer Night's Dream* (*c.* 1595), one of the most successful of all his plays, displays the kind of multiple plotting he had practiced in *The Taming of the Shrew* and other earlier comedies. The overarching plot is of Duke Theseus of Athens and his impending marriage to an Amazonian warrior, Hippolyta, whom Theseus has recently conquered and taken back to Athens to be his bride. Their marriage ends the play. They share this concluding ceremony with the four young lovers Hermia and Lysander and Helena and Demetrius, who have fled into the forest nearby to escape the Athenian law and to pursue one another, whereupon they are subjected to a complicated series of mix-ups. Eventually all is righted by fairy magic, though the fairies are no less at strife. Oberon, king of the fairies, quarrels with his Queen Titania over a changeling boy and punishes her by causing her to fall in love with an Athenian artisan who wears an ass's head. The artisans are in the forest to rehearse a play for the forthcoming marriage of Theseus and Hippolyta. Thus, four separate strands or plots interact with one another. Despite the play's brevity, it is a masterpiece of artful construction.

The use of multiple plots encourages a varied treatment of the experiencing of love. For the two young human couples, falling in love is quite hazardous; the long-standing friendship between the two young women is threatened and almost destroyed by the rivalries of heterosexual encounter. The eventual transition to heterosexual marriage seems to them to have been a process of dreaming, indeed of nightmare, from which they emerge miraculously restored to their best selves. Meantime, the marital strife of Oberon and Titania is, more disturbingly, one in which the female is humiliated until she submits to the will of her husband. Similarly, Hippolyta is an Amazon warrior queen who has had to submit to the authority of a husband. Fathers and daughters are no less at strife until, as in a dream, all is resolved by the magic of Puck and Oberon. Love is ambivalently both an enduring ideal relationship and a struggle for mastery in which the male has the upper hand.

Conflict between the sexes in *A Midsummer Night's Dream*

The Merchant of Venice (*c.* 1596–97) uses a double plot structure to contrast a tale of romantic wooing with one that comes close to tragedy. Portia is a fine example of a romantic heroine in Shakespeare's mature comedies: she is witty, rich, exacting in what she expects of men, and adept at putting herself in a male disguise to make her presence felt. She is loyally obedient to her father's will and yet determined that she shall have Bassanio. She triumphantly resolves the murky legal affairs of Venice when the men have all failed. Shylock, the Jewish moneylender, is at the point of exacting a pound of flesh from Bassanio's friend Antonio as payment for a forfeited loan. Portia foils him in his attempt in a way that is both clever and shystering. Sympathy is uneasily balanced in Shakespeare's portrayal of Shylock, who is both persecuted by his Christian opponents and all too ready to demand an eye for an eye according to Old Testament laws. Ultimately Portia triumphs, not only with Shylock in the court of law but in her marriage with Bassanio.

Much Ado About Nothing (*c.* 1598–99) revisits the issue of power struggles in courtship, again in a revealingly double plot. The young heroine of the more conventional story, derived from Italianate fiction, is wooed by a respectable young aristocrat named Claudio who has won his spurs and now considers it his pleasant duty to take a wife. He knows so little about Hero (as she is named) that he gullibly credits the contrived evidence of the play's villain, Don John, that she has had many lovers, including one on the evening before the intended wedding. Other men as well, including Claudio's senior officer, Don Pedro, and Hero's father, Leonato, are all too ready to believe the slanderous accusation. Only comic circumstances rescue Hero from her accusers and reveal to the men that they have been fools. Meantime, Hero's cousin, Beatrice, finds it hard to overcome her skepticism about men, even when she is wooed by Benedick, who is also a skeptic about marriage. Here the barriers to romantic understanding are inner and psychological and must be defeated by the good-natured plotting of their friends. In what could be regarded a brilliant rewriting of *The Taming of the Shrew*, the witty battle of the sexes is no less amusing and complicated, but the eventual accommodation finds something much closer to mutual respect and equality between men and women.

Rosalind, in *As You Like It* (*c.* 1598–1600), makes use of the by-now-familiar device of disguise as a young man in order to pursue the ends of promoting a rich and substantial relationship between the sexes. As in other of these plays, Rosalind is more emotionally stable and mature than her young man, Orlando. He lacks formal education and is all rough edges, though fundamentally decent and attractive. She is the daughter of the banished Duke who finds herself obliged, in turn, to go into banishment with her dear cousin Celia and the court fool, Touchstone. Although Rosalind's male disguise is at first a means of survival in a seemingly inhospitable forest, it soon serves a more interesting function. As "Ganymede," Rosalind befriends Orlando, offering him counseling in the affairs of love. Orlando, much in need of such advice, readily accepts and proceeds to woo his "Rosalind" ("Ganymede" playing her own self) as though she were indeed a woman. Her wryly amusing perspectives on the follies of young love helpfully puncture Orlando's inflated and unrealistic "Petrarchan" stance as the young lover who writes poems to his mistress and sticks them up on trees. Other figures in the play further an understanding of love's glorious foolishness by their various attitudes: Silvius, the pale-faced wooer out of pastoral romance; Phoebe, the disdainful mistress whom he worships; William, the country bumpkin, and Audrey, the country wench; and, surveying and commenting on every imaginable kind of human folly, the clown Touchstone and the malcontent traveler Jaques.

Male impersonation in *As You Like It*

Twelfth Night (1600–02) pursues a similar motif of female disguise. Viola, cast ashore in Illyria by a shipwreck and obliged to disguise herself as a young man in order to gain a place in the court of Duke Orsino, falls in love with the duke and uses her disguise as a cover for an educational process not unlike that given by Rosalind to Orlando. Orsino is as unrealistic a lover as one could hope to imagine; he pays fruitless court to the Countess Olivia and seems content with the unproductive love melancholy in which he wallows. Only Viola, as "Cesario," is able to awaken in him a genuine feeling for friendship and love. They become inseparable companions and then seeming rivals for the hand of Olivia until the presto-change of Shakespeare's stage magic is able to restore "Cesario" to her woman's garments and thus present to Orsino the flesh-and-blood woman that he has only distantly imagined. The transition from same-sex friendship to heterosexual union is a constant in Shakespearean comedy. The woman is the self-knowing, constant, loyal one; the man needs to learn a lot from the woman. As in the other plays as well, *Twelfth Night* neatly plays off this courtship theme with a second plot, of Malvolio's self-deception that he is desired by Olivia—an illusion that can be addressed only by the satiric devices of exposure and humiliation.

The qualities of women in *Twelfth Night*

The Merry Wives of Windsor (1597–1601) is an interesting deviation from the usual Shakespearean romantic comedy in that it is set not in some imagined far-off place like Illyria or Belmont or the forest of Athens but in Windsor, a solidly bourgeois village near Windsor Castle in the heart of England. Uncertain tradition has it that Queen Elizabeth wanted to see Falstaff in love. There is little, however, in the way of romantic wooing (the story of Anne Page and her suitor Fenton is rather buried in the

midst of so many other goings-on), but the play's portrayal of women, and especially of the two "merry wives," Mistress Alice Ford and Mistress Margaret Page, reaffirms what is so often true of women in these early plays, that they are good-hearted, chastely loyal, and wittily self-possessed. Falstaff, a suitable butt for their cleverness, is a scapegoat figure who must be publicly humiliated as a way of transferring onto him the human frailties that Windsor society wishes to expunge.

Falstaff as a scapegoat

Completion of the histories. Concurrent with his writing of these fine romantic comedies, Shakespeare also brought to completion (for the time being at least) his project of writing 15th-century English history. After having finished in 1589–94 the tetralogy about Henry VI, Edward IV, and Richard III, bringing the story down to 1485, and then a play about *King John* (*c.* 1594–96) that deals with a chronological period (the 13th century) that sets it quite apart from his other history plays, Shakespeare returned to the late 14th and early 15th centuries and to the chronicle of Richard II, Henry IV, and Henry's legendary son Henry V. This inversion of historical order in the two tetralogies allowed Shakespeare to finish his sweep of late medieval English history with Henry V, a hero king in a way that Richard III could never pretend to be.

Richard II (*c.* 1595–96), written throughout in blank verse, is a sombre play about political impasse. Richard, installed at an early age into the kingship, proves irresponsible as a ruler. He unfairly banishes his own first cousin, Henry Bolingbroke (later to be Henry IV). When Richard keeps the dukedom of Lancaster from Bolingbroke without proper legal authority, he manages to alienate many nobles and to encourage Bolingbroke's return from exile. That return, too, is illegal, but it is a fact, and when several of the nobles (including York) go over to Bolingbroke's side, Richard is forced to abdicate. The rights and wrongs of this power struggle are masterfully ambiguous. History proceeds without any sense of moral imperative. Henry IV is a more capable ruler, but his authority is tarnished by his crimes, and his own rebellion appears to teach the barons to rebel against him in turn. Henry eventually dies a disappointed man.

The dying King Henry IV must turn royal authority over to young Hal, or Henry, now Henry V. The prospect is dismal both to the dying king and to the members of his court, for Prince Hal has distinguished himself to this point mainly by his penchant for keeping company with the disreputable if engaging Falstaff. The son's attempts at reconciliation with the father succeed temporarily, especially when Hal saves his father's life at the battle of Shrewsbury, but (especially in *2 Henry IV*) his reputation as a wastrel will not leave him. Everyone expects from him a reign of irresponsible license, with Falstaff in an influential position. It is for these reasons that the young king must publicly repudiate his old companion of the tavern and the highway, however much that repudiation tugs at his heart and the audience's. Falstaff, for all his debauchery and irresponsibility, is infectiously amusing and delightful; he represents in Hal a spirit of youthful vitality that is left behind only with the greatest of regret as the young man assumes manhood and the role of crown prince. Hal manages all this with aplomb and goes on to defeat the French mightily at the battle of Agincourt. Even his high jinks are a part of what is so attractive in him. Maturity and position come at a great personal cost: Hal becomes less a frail human being and more the figure of royal authority.

Young Hal and Falstaff

Thus, in his plays of the 1590s, the young Shakespeare concentrated to a remarkable extent on romantic comedies and English history plays. The two genres are nicely complementary: the one deals with courtship and marriage, while the other examines the career of a young man growing up to be a worthy king. Only at the end of the history plays does Henry V have any kind of romantic relationship with a woman, and this one instance is quite unlike courtships in the romantic comedies: Hal is given the Princess of France as his prize, his reward for sturdy manhood. He takes the lead in the wooing scene in which he invites her to join him in a political marriage. In both romantic comedies and English history plays, a young man successfully negotiates the hazardous and potentially rewarding paths of sexual and social maturation.

Romeo and Juliet. Apart from the early *Titus Andronicus,* the only other play that Shakespeare wrote prior to 1599 that is classified as a tragedy is *Romeo and Juliet* (1594–96), which is quite untypical of the tragedies that were to follow. Written more or less at the time when Shakespeare was writing *A Midsummer Night's Dream, Romeo and Juliet* shares many of the characteristics of romantic comedy. Romeo and Juliet are not persons of extraordinary social rank or position, like Hamlet, Othello, King Lear, and Macbeth. They are the boy and girl next door, interesting not for their philosophical ideas but for their appealing love for each other. They are character types more suited to classical comedy in that they do not derive from the upper class. Their wealthy families are essentially bourgeois. The eagerness with which Capulet and his wife court Count Paris as their prospective son-in-law bespeaks their desire for social advancement.

Romeo and Juliet as bourgeois lovers

Accordingly, the first half of *Romeo and Juliet* is very funny, while its delight in verse forms reminds us of *A Midsummer Night's Dream.* The bawdry of Mercutio and of the Nurse is richly suited to the comic texture of the opening scenes. Romeo, haplessly in love with a Rosaline whom we never meet, is a partly comic figure like Silvius in *As You Like It.* The plucky and self-knowing Juliet is much like the heroines of romantic comedies. She is able to instruct Romeo in the ways of speaking candidly and unaffectedly about their love rather than in the frayed cadences of the Petrarchan wooer.

The play is ultimately a tragedy, of course, and indeed warns its audience at the start that the lovers are "star-crossed." Yet the tragic vision is not remotely that of *Hamlet* or *King Lear.* Romeo and Juliet are unremarkable, nice young people doomed by a host of considerations outside themselves: the enmity of their two families, the misunderstandings that prevent Juliet from being able to tell her parents whom it is that she has married, and even unfortunate coincidence (such as the misdirection of the letter sent to Romeo to warn him of the Friar's plan for Juliet's recovery from a deathlike sleep). Yet there is the element of personal responsibility upon which most mature tragedy rests when Romeo chooses to avenge the death of Mercutio by killing Tybalt, knowing that this deed will undo the soft graces of forbearance that Juliet has taught him. Romeo succumbs to the macho peer pressure of his male companions, and tragedy results in part from this choice. Yet so much is at work that the reader ultimately sees *Romeo and Juliet* as a love tragedy—celebrating the exquisite brevity of young love, regretting an unfeeling world, and evoking an emotional response that differs from that produced by the tragedies. Romeo and Juliet are, at last, "Poor sacrifices of our enmity" (5.3.304). The emotional response the play evokes is a strong one, but it is not like the response called forth by the tragedies after 1599.

The "problem" plays. Whatever his reasons, about 1599–1600 Shakespeare turned with unsparing intensity to the exploration of darker issues such as revenge, sexual jealousy, aging, midlife crisis, and death. Perhaps he saw that his own life was moving into a new phase of more complex and vexing experiences. Perhaps he felt, or sensed, that he had worked through the romantic comedy and history play and the emotional trajectories of maturation that they encompassed. In any event, he began writing not only his great tragedies but a group of plays that are hard to classify in terms of genre. They are sometimes grouped today as "problem" plays or "problem" comedies. An examination of these plays is crucial to understanding this period of transition from about 1599 to 1603.

The three problem plays dating from these years are *All's Well That Ends Well, Measure for Measure,* and *Troilus and Cressida. All's Well* is a comedy ending in acceptance of marriage, but in a way that poses thorny ethical issues. Count Bertram cannot initially accept his marriage to Helena, a woman of lower social station who has grown up in his noble household and has won Bertram as her husband by her seemingly miraculous cure of the French

king. Bertram's reluctance to face the responsibilities of marriage is all the more dismaying when he turns his amorous intentions to a Florentine maiden, Diana, whom he wishes to seduce without marriage. Helena's stratagem to resolve this difficulty is the so-called "bed trick," substituting herself in Bertram's bed for the arranged assignation and then calling her wayward husband to account when she is pregnant with his child. Her ends are achieved by such morally ambiguous means that marriage seems at best a precarious institution on which to base the presumed reassurances of romantic comedy.

The bed trick in *Measure for Measure*

Measure for Measure similarly employs the bed trick, and for a similar purpose, though in even murkier circumstances. Isabella, on the verge of becoming a nun, learns that she has attracted the sexual desire of Lord Angelo, the deputy ruler of Vienna serving in the mysterious absence of the Duke. Her plea to Angelo for her brother's life, when that brother (Claudio) has been sentenced to die for fornication with his fiancée, is met with a demand that she sleep with Angelo or forfeit Claudio's life. This ethical dilemma is resolved by a trick (devised by the Duke, in disguise) to substitute for Isabella a woman (Mariana) whom Angelo was supposed to marry but has refused when she can produce no dowry. The Duke's motivations in manipulating these substitutions and false appearances are unclear, though arguably his wish is to see what the various characters of this play will do when faced with seemingly impossible choices. Angelo is revealed as a morally fallen man, a would-be seducer and murderer who is nonetheless remorseful and ultimately glad to have been prevented from carrying out his intended crimes; Claudio learns that he is coward enough to wish to live by any means, including the emotional and physical blackmail of his sister; and Isabella learns that she is capable of bitterness and hatred, even if, crucially, she finally discovers that she can and must forgive her enemy. Her charity, and the Duke's stratagems, make possible an ending in forgiveness and marriage, but in that process the nature and meaning of marriage are severely tested.

Troilus and Cressida (*c.* 1601) is the most experimental and puzzling of these three plays. Simply in terms of genre, it is virtually unclassifiable. It can hardly be a comedy, ending as it does in the deaths of Patroclus and Hector and the looming defeat of the Trojans. Nor is the ending normative in terms of romantic comedy: the lovers, Troilus and Cressida, are separated from one another and embittered by the failure of their relationship. The play is a history play in a sense, dealing as it does with the great Trojan war celebrated in Homer's *Iliad,* and yet its purpose is hardly that of telling the story of the war. As a tragedy it is perplexing in that the chief figures of the play (apart from Hector) do not die at the end, and the mood is one of desolation and even disgust rather than tragic catharsis. Perhaps the play should be thought of as a satire; the choric observations of Thersites and Pandarus serve throughout as a mordant commentary on the interconnectedness of war and lechery. With fitting ambiguity, the play was placed in the Folio of 1623 between the histories and the tragedies, in a category all by itself. Clearly, in these problem plays Shakespeare was opening up for himself a host of new problems in terms of genre and human sexuality.

Julius Caesar. Written in 1599 (the same year as *Henry V*), probably for the opening of the Globe Theatre on the south bank of the Thames, *Julius Caesar* illustrates similarly the transition in Shakespeare's writing toward darker themes and tragedy. It, too, is a history play in a sense, dealing with a non-Christian civilization existing 16 centuries before Shakespeare wrote his plays. Roman history opened up for Shakespeare a world in which divine purpose could not be easily ascertained. The characters of *Julius Caesar* variously interpret the great event of the assassination of Caesar as one in which the gods are angry, or disinterested, or capricious, or simply not there. The wise Cicero observes, "Men may construe things after their fashion, / Clean from the purpose of the things themselves" (1.3.34–35).

The obscurity of divine purpose in *Julius Caesar*

Human history in *Julius Caesar* seems to follow an undular pattern of rise and fall, in a way that is cyclical rather than divinely purposeful. Caesar enjoys his days of triumph, until he is cut down by the conspirators; Brutus and Cassius succeed to power, but not for long. Brutus's attempts to protect Roman republicanism and the freedom of the city's citizens to govern themselves through senatorial tradition ends up in the destruction of the very liberties he most cherished. He and Cassius meet their destiny at the battle of Philippi. They are truly tragic figures, especially Brutus, in that their essential characters are their fate; Brutus is a good man but is also proud and stubborn, and these latter qualities ultimately bring about his death. Shakespeare's first major tragedy is Roman in spirit and classical in its notion of tragic character. It shows what Shakespeare had to learn from classical precedent as he set about looking for workable models in tragedy.

The tragedies. *Hamlet* (*c.* 1599–1601), on the other hand, chooses a tragic model closer to that of *Titus Andronicus* and Kyd's *The Spanish Tragedy.* In form, *Hamlet* is a revenge tragedy. It features characteristics found in *Titus* as well: a protagonist charged with the responsibility of avenging a heinous crime against the protagonist's family; a cunning antagonist; the appearance of the ghost of the murdered person; the feigning of madness to throw off the villain's suspicions; the play within a play as a means of testing the villain; and still more.

Yet to search out these comparisons is to highlight what is so extraordinary about *Hamlet,* for it refuses to be merely a revenge tragedy. Shakespeare's protagonist is unique in the genre in his moral qualms and most of all in his finding a way to carry out his dread command without becoming a cold-blooded murderer. Hamlet does act bloodily, especially when he kills Polonius, thinking that the old man hidden in Gertrude's chambers must be the King, whom Hamlet is commissioned to kill. The act seems plausible and strongly motivated, and yet Hamlet sees at once that he has erred. He has killed the wrong man, even if Polonius has brought this on himself with his incessant spying. Hamlet sees that he has offended heaven and that he will have to pay for his act. When, at the play's end, Hamlet encounters his fate in a duel with Polonius's son, Laertes, Hamlet interprets his own tragic story as one that providence has made meaningful. By placing himself in the hands of providence and believing devoutly that "There's a divinity that shapes our ends, / Rough-hew them how we will" (5.2.10–11), Hamlet finds himself ready for a death that he has longed for. He also finds an opportunity for killing Claudius almost unpremeditatedly, spontaneously, as an act of reprisal for all that Claudius has done.

Hamlet interprets his fate

Hamlet thus finds tragic meaning in his own story. More broadly, too, he has searched for meaning in dilemmas of all sorts: his mother's overhasty marriage, Ophelia's weak-willed succumbing to the will of her father and brother, his being spied on by his erstwhile friends Rosencrantz and Guildenstern, and much more. His utterances are often despondent, relentlessly honest, and philosophically profound as he ponders the nature of friendship, memory, romantic attachment, filial love, sensuous enslavement, corrupting habits (drinking, sexual lust), and almost every phase of human experience.

One remarkable aspect about Shakespeare's great tragedies (*Hamlet, Othello, King Lear, Macbeth,* and *Antony and Cleopatra* most of all) is that they proceed through such a staggering range of human emotions, and especially the emotions that are appropriate to the mature years of the human cycle. Hamlet is 30, one learns—an age when a person is apt to perceive that the world around him is "an unweeded garden / That grows to seed. Things rank and gross in nature / Possess it merely" (1.2.135–137). Shakespeare was about 36 when he wrote this play. *Othello* (*c.* 1603–04) centres on sexual jealousy in marriage. *King Lear* (*c.* 1605–06) is about aging, generational conflict, and feelings of ingratitude. *Macbeth* (*c.* 1606–07) explores mad ambition. *Antony and Cleopatra,* written about 1606–07, when Shakespeare was 42 or thereabouts, studies the exhilarating but ultimately dismaying phenomenon of midlife crisis.

These plays are deeply concerned with domestic and

family relationships. In *Othello* Desdemona is the only daughter of Brabantio, an aging senator of Venice, who dies heartbroken because his daughter has eloped with a man who is her senior by many years and is of another race. With Othello, Desdemona is briefly happy, despite her filial disobedience, until a terrible sexual jealousy is awakened in him, quite without cause other than his own fears and susceptibility to Iago's insinuations that it is only "natural" for Desdemona to seek erotic pleasure with a young man of her own race. Driven by his own deeply irrational fear and hatred of women and seemingly mistrustful of his own masculinity, Iago can assuage his own inner hell only by persuading other men like Othello that their inevitable fate is to be cuckolded. As a tragedy, the play adroitly exemplifies the traditional classical model of a good man brought to misfortune by *hamartia,* or tragic flaw; as Othello grieves, he is one who has "loved not wisely, but too well" (5.2.354). It bears remembering, however, that Shakespeare owed no loyalty to this classical model; *Hamlet,* for one, is a play that does not work well in Aristotelian terms.

Daughters and fathers are also at the heart of the major dilemma in *King Lear.* In this configuration, Shakespeare does what he often does in his late plays: erase the wife from the picture, so that father and daughter(s) are left to deal with one another. (Compare *Othello, The Winter's Tale, Cymbeline, The Tempest,* and perhaps the circumstances of Shakespeare's own life, in which his relations with his daughter Susanna especially seem to have meant more to him than his partly estranged marriage with Anne.) Lear's banishing of his favourite daughter, Cordelia, because of her laconic refusal to proclaim a love for him as the essence of her being, brings upon this aging king the terrible punishment of being belittled and rejected by his ungrateful daughters, Goneril and Regan. Concurrently, in the play's second plot, the Earl of Gloucester makes a similar mistake with his good-hearted son, Edgar, and thereby delivers himself into the hands of his scheming bastard son, Edmund. Both these erring elderly fathers are ultimately nurtured by the loyal children they have banished, but not before the play has tested to its absolute limit the proposition that evil can flourish in a bad world.

The indifferent gods in *King Lear*

The gods seem indifferent, perhaps absent entirely; pleas to them for assistance go unheeded while the storm of fortune rains down on the heads of those who have trusted in conventional pieties. Part of what is so great in this play is that its testing of the major characters requires them to seek out philosophical answers that can arm the resolute heart against ingratitude and misfortune by constantly pointing out that life owes one nothing. The consolations of philosophy preciously found out by Edgar and Cordelia are those that rely not on the suppositious gods but on an inner moral strength demanding that one be charitable and honest because life is otherwise monstrous and subhuman. The play exacts terrible prices of those who persevere in goodness, but it leaves them and the reader, or audience, with the reassurance that it is simply better to be a Cordelia than to be a Goneril, to be an Edgar than to be an Edmund.

Macbeth is in some ways Shakespeare's most unsettling tragedy because it invites the intense examination of the heart of a man who is well-intentioned in most ways but discovers he cannot resist the temptation to achieve power at any cost. Macbeth is a sensitive, even poetic person, and as such he understands with frightening clarity the stakes involved in his contemplated deed of murder. Duncan is a virtuous king and his guest. The murder is a violation of the sacred obligations of hospitality. Macbeth knows that Duncan's virtues, like angels, "trumpet-tongued," will plead against "the deep damnation of his taking-off" (1.7.19–20). The question of why he proceeds to murder is partly answered by the insidious temptations of the three Weird Sisters and the terrifying strength of his wife, who drives him on to murder by describing his reluctance as unmanliness. Ultimately, though, the responsibility lies with Macbeth. His collapse of moral integrity confronts the audience and perhaps implicates it. The loyalty and decency of such characters as Macduff hardly offset what is so painfully weak in the play's protagonist.

Antony and Cleopatra approaches human frailty in terms that are less spiritually terrifying. The story of the lovers is certainly one of worldly failure. Plutarch's *Lives* gave to Shakespeare the object lesson of a brave general who lost his reputation and sense of self-worth through his infatuation with an admittedly attractive but nonetheless dangerous woman. Shakespeare changes none of the circumstances: Antony hates himself for dallying in Egypt with Cleopatra, agrees to marry with Octavius Caesar's sister Octavia as a way of recovering his status in the Roman triumvirate, cheats on Octavia eventually, loses the battle of Actium because of his fatal attraction for Cleopatra, and dies in Egypt a defeated, aging warrior. Shakespeare adds to this narrative a compelling portrait of midlife crisis. Antony is deeply anxious about his loss of sexual potency and position in the world of affairs. His amorous life in Egypt is manifestly an attempt to affirm and recover his dwindling male power.

Midlife crisis in *Antony and Cleopatra*

Yet the Roman model is not in Shakespeare's play the unassailably virtuous choice that it is in Plutarch. In *Antony and Cleopatra* Roman behaviour does promote attentiveness to duty and worldly achievement, but, as embodied in young Octavius, it is also obsessively male and cynical about women. Octavius is intent on capturing Cleopatra and leading her in triumph back to Rome—that is, to cage the unruly woman and place her under male control. When Cleopatra perceives that aim, she chooses a noble suicide rather than humiliation by a patriarchal male. In her suicide, Cleopatra avers that she has called "great Caesar ass / Unpolicied" (5.2.307–308). Vastly to be preferred is the fleeting dream of greatness with Antony, both of them unfettered, godlike, like Isis and Osiris, immortalized as heroic lovers even if the actual circumstances of their lives were often disappointing and even tawdry. The vision in this tragedy is deliberately unstable, but at its most ethereal it encourages a vision of human greatness that is distant from the soul-corrupting evil of *Macbeth* or *King Lear.*

Two late tragedies also choose the ancient classical world as their setting but do so in a deeply dispiriting way. Shakespeare appears to have been much preoccupied with ingratitude and human greed in these years. *Timon of Athens* (*c.* 1605–08), probably an unfinished play and possibly never produced, initially reveals a prosperous man fabled for his generosity. When he discovers that he has exceeded his means, he turns to his seeming friends for the kinds of assistance he has given them, only to discover that their memories are short. Retiring to a bitter isolation, Timon rails against all humanity and refuses every sort of consolation, even that of well-meant companionship and sympathy from a former servant. He dies in isolation. The unrelieved bitterness of this account is only partly ameliorated by the story of the military captain Alcibiades, who has also been the subject of Athenian ingratitude and forgetfulness but manages to reassert his authority at the end. Alcibiades resolves to make some accommodation with the wretched condition of humanity; Timon will have none of it. Seldom has a more unrelievedly embittered play been written.

Coriolanus (*c.* 1608) similarly portrays the ungrateful responses of a city toward its military hero. The problem is complicated by the fact that Coriolanus, egged on by his mother and his conservative allies, undertakes a political role in Rome for which he is not temperamentally fitted. His friends urge him to hold off his intemperate speech until he is voted into office, but Coriolanus is too plain-spoken to be tactful in this way. His contempt for the plebeians and their political leaders, the tribunes, is unsparing. His political philosophy, while relentlessly aristocratic and snobbish, is consistent and theoretically sophisticated; the citizens are, as he argues, incapable of governing themselves judiciously. Yet his fury only makes matters worse and leads to an exile from which he returns to conquer his own city, in league with his old enemy and friend, Aufidius. When his mother comes out for the city to plead for her life and that of other Romans, he relents and thereupon falls into defeat as a kind of mother's boy, unable to assert his own sense of self. As a tragedy, *Coriolanus* is again bitter and satiric, ending in defeat and hu-

Self-destructiveness in *Coriolanus*

miliation. It is an immensely powerful play, and it captures a philosophical mood of nihilism and bitterness that hovers over Shakespeare's writings throughout these years in the first decade of the 1600s.

The romances. Concurrently, nonetheless, and then in the years that followed, Shakespeare turned again to the writing of comedy. The late comedies are usually called romances, or tragicomedies, because they tell stories of wandering and separation leading eventually to tearful and joyous reunion. They are suffused with a bittersweet mood that seems eloquently appropriate to a writer who has explored with such unsparing honesty the depths of human suffering and degradation in the great tragedies.

Bittersweet mood of the late romances

Pericles, written perhaps in 1606–08 and based on the familiar tale of Apollonius of Tyre, may involve some collaboration of authorship; the text is unusually imperfect and did not appear in the Folio of 1623. It employs a chorus figure, John Gower (author of an earlier version of this story), to guide the reader or viewer around the Mediterranean on Pericles' various travels as he avoids marriage with the daughter of the incestuous King Antiochus of Antioch; marries Thaisa, the daughter of King Simonides of Pentapolis; has a child by her; believes his wife to have died in childbirth during a storm at sea and has her body thrown overboard to quiet the superstitious fears of the sailors; puts his daughter Marina in the care of Cleon of Tarsus and his wicked wife, Dionyza; and is eventually restored to his wife and child after many years. The story is typical romance. Shakespeare adds touching scenes of reunion and a perception that beneath the naive account of travel lies a subtle dramatization of separation, loss, and recovery. Pericles is deeply burdened by his loss and perhaps too a sense of guilt for having consented to consigning his wife's body to the sea. He is recovered from his despair only by the ministrations of a loving daughter, who is able to give him a reason to live again and then to be reunited with his wife.

The Winter's Tale (*c.* 1609–11) is in some ways a replaying of this same story in that King Leontes of Sicilia, smitten by an irrational jealousy of his wife, Hermione, brings about the seeming death of that wife and the real death of their son. The resulting guilt is unbearable for Leontes and yet is ultimately curative over the period of many years required for his only daughter, Perdita (whom he has nearly killed also), to grow to maturity in distant Bohemia. This story too is based on a prose romance, in this case Robert Greene's *Pandosto.* The reunion with daughter and then wife is deeply touching as in *Pericles,* with the added magical touch that the audience does not know that Hermione is alive and in fact has been told that she is dead. Her wonderfully staged appearance as a statue coming to life is one of the great theatrical coups in Shakespeare, playing as it does with favourite Shakespearean themes in these late plays of the ministering daughter, the guilt-ridden husband, and the miraculously recovered wife. The story is all the more moving when one considers that Shakespeare may have had, or imagined, a similar experience of attempting to recover a relationship with his wife, Anne, whom he had left in Stratford during his many years in London.

Leontes' reunion with Hermione in *The Winter's Tale*

In *Cymbeline* (*c.* 1608–10) King Cymbeline drives his virtuous daughter, Imogen, into exile by his opposition to her marriage with Posthumus Leonatus. The wife in this case is Cymbeline's baleful Queen, a stereotypical wicked stepmother whose witless and lecherous son, Cloten (Imogen's half-brother), is the embodiment of everything that threatens and postpones the eventual happy ending of this tale. Posthumus too fails Imogen by being irrationally jealous of her, but he is eventually recovered to a belief in her goodness. The dark portraiture of the queen illustrates how ambivalent Shakespeare's view of the mother is in his late plays. This Queen is the wicked stepmother, like Dionyza in *Pericles;* in her relentless desire for control, she also brings to mind Lady Macbeth and the Weird Sisters in *Macbeth,* as well as Coriolanus's mother, Volumnia. The devouring mother is a forbidding presence in the late plays, though she is counterbalanced by redeeming maternal figures such as Hermione in *The Winter's Tale* and Thaisa in *Pericles.*

The Tempest (*c.* 1611) sums up much of what Shakespeare's mature art was all about. Once again we find a wifeless father with a daughter, in this case on a deserted island where the father, Prospero, is entirely responsible for his daughter's education. He behaves like a dramatist in charge of the whole play as well, arranging her life and that of the other characters. He employs a storm at sea to bring young Ferdinand into the company of his daughter; Ferdinand is Prospero's choice because such a marriage will resolve the bitter dispute between Milan and Naples—which arose after the latter supported Prospero's usurping brother, Antonio, in his claim to the dukedom of Milan—that has led to Prospero's banishment. At the same time, Ferdinand is certainly Miranda's choice as well; the two fall instantly in love, anticipating the desired romantic happy ending. The ending will also mean an end to Prospero's career as artist and dramatist, for he is nearing retirement and senses that his gift will not stay with him forever. The imprisoned spirit Ariel, embodiment of that temporary and precious gift, must be freed in the play's closing moments. Caliban, too, must be freed, since Prospero has done what he can to educate and civilize this Natural Man. Art can go only so far.

The Tempest seems to have been intended as Shakespeare's farewell to the theatre. It contains moving passages of reflection on what his powers as artist have been able to accomplish, and valedictory themes of closure. As a comedy, it demonstrates perfectly the way that Shakespeare was able to combine precise artistic construction (the play chooses on this farewell occasion to observe the classical unities of time, place, and action) with his special flair for stories that transcend the merely human and physical: *The Tempest* is peopled with spirits, monsters, and drolleries. Here, it seems, is Shakespeare's summation of his art as comic dramatist.

The Tempest as Shakespeare's farewell to theatre

But *The Tempest* proved not to be Shakespeare's last play after all. Perhaps he discovered, as many people do, that he was bored in retirement in 1613 or thereabouts. No doubt his acting company was eager to have him back. He wrote a history play titled *Henry VIII* (1613), which is extraordinary in a number of ways: it relates historical events substantially later chronologically than those of the 15th century that had been his subject in his earlier historical plays; it is separated from the last of those plays by perhaps 14 years; and, perhaps most significant, it is as much romance as history play. History in this instance is really about the birth of Elizabeth I, who was to become England's great queen. The circumstances of Henry VIII's troubled marital affairs, his meeting with Anne Boleyn, his confrontation with the papacy, and all the rest turn out to be the humanly unpredictable ways by which providence engineers the miracle of Elizabeth's birth. The play ends with this great event and sees in it a justification and necessity of all that has proceeded. Thus, history yields its providential meaning in the shape of a play that is both history and romance.

Collaborations and spurious attributions. *The Two Noble Kinsmen* (1613–14) brought Shakespeare into collaboration with John Fletcher, his successor as chief playwright for the King's Men. (Fletcher is sometimes thought also to have helped Shakespeare with *Henry VIII.*) The story, taken out of Chaucer's *Knight's Tale,* is essentially another romance, in which two young gallants compete for the hand of Emilia and in which deities preside over the choice. Shakespeare may have had a hand earlier as well in *Edward III,* a history play of about 1590–95, and he seems to have provided a scene or so for *The Book of Sir Thomas More* (*c.* 1593–1601) when that play encountered trouble with the censor. Collaborative writing was common in the Renaissance English stage, and it is not surprising that Shakespeare was called upon to do some of it. Nor is it surprising that, given his towering reputation, he was credited with having written a number of plays that he had nothing to do with, including those that were spuriously added to the third edition of the Folio in 1664: *Locrine* (1591–95), *Sir John Oldcastle* (1599–1600), *Thomas Lord Cromwell* (1599–1602), *The London Prodigal* (1603–05), *The Puritan* (1606), and *A Yorkshire Tragedy* (1605–08). To a remarkable extent, nonetheless,

his corpus stands as a coherent body of his own work. The shape of the career has a symmetry and internal beauty not unlike that of the individual plays and poems.

(D.Bev.)

Shakespeare's reading

With a few exceptions, Shakespeare did not invent the plots of his plays. Sometimes he used old stories (*Hamlet, Pericles*). Sometimes he worked from the stories of comparatively recent Italian writers, such as Boccaccio—using both well-known stories (*Romeo and Juliet, Much Ado About Nothing*) and little-known ones (*Othello*). He used the popular prose fictions of his contemporaries in *As You Like It* and *The Winter's Tale.* In writing his historical plays, he drew largely from Sir Thomas North's translation of Plutarch's *Lives of the Noble Grecians and Romans* for the Roman plays and the chronicles of Edward Hall and Ralph Holinshed for the plays based upon English history. Some plays deal with rather remote and legendary history (*King Lear, Cymbeline, Macbeth*). Earlier dramatists had occasionally used the same material (there were, for example, the earlier plays called *The Famous Victories of Henry the Fifth* and *King Leir*). But, because many plays of Shakespeare's time have been lost, it is impossible to be sure of the relation between an earlier, lost play and Shakespeare's surviving one: in the case of *Hamlet* it has been plausibly argued that an "old play," known to have existed, was merely an early version of Shakespeare's own.

Shakespeare was probably too busy for prolonged study. He had to read what books he could, when he needed them. His enormous vocabulary could be derived only from a mind of great celerity, responding to the literary as well as the spoken language. It is not known what libraries were available to him. The Huguenot family of Mountjoys, with whom he lodged in London, presumably possessed French books. Moreover, he seems to have enjoyed an interesting connection with the London book trade. The Richard Field who published Shakespeare's two poems *Venus and Adonis* and *The Rape of Lucrece,* in 1593–94, seems to have been (as an apprenticeship record describes him) the "son of Henry Field of Stratford-upon-Avon in the County of Warwick, tanner." When Henry Field the tanner died in 1592, John Shakespeare the glover was one of the three appointed to value his goods and chattels. Field's son, bound apprentice in 1579, was probably about the same age as Shakespeare. From 1587 he steadily established himself as a printer of serious literature—notably of North's translation of Plutarch (1595, reprinted in 1603 and 1610). There is no direct evidence of any close friendship between Field and Shakespeare. Still, it cannot escape notice that one of the important printer-publishers in London at the time was an exact contemporary of Shakespeare at Stratford, that he can hardly have been other than a schoolfellow, that he was the son of a close associate of John Shakespeare, and that he published Shakespeare's first poems. Clearly, a considerable number of literary contacts were available to Shakespeare, and many books were accessible.

Shakespeare's acquaintance with London printer

That Shakespeare's plays had "sources" was already apparent in his own time. An interesting contemporary description of a performance is to be found in the diary of a young lawyer of the Middle Temple, John Manningham, who kept a record of his experiences in 1602 and 1603. On February 2, 1602, he wrote:

> At our feast we had a play called *Twelfth Night,* or *What You Will,* much like *The Comedy of Errors,* or *Menaechmi* in Plautus, but most like and near to that in Italian called *Inganni.*

The first collection of information about sources of Elizabethan plays was published in the 17th century—Gerard Langbaine's *Account of the English Dramatick Poets* (1691) briefly indicated where Shakespeare found materials for some plays. But, during the course of the 17th century, it came to be felt that Shakespeare was an outstandingly "natural" writer, whose intellectual background was of comparatively little significance: "He was naturally learn'd; he needed not the spectacles of books to read nature," wrote John Dryden in 1668.

The first collection of source materials, arranged so that they could be read and closely compared with Shakespeare's plays, was made by Mrs. Charlotte Lennox in the 18th century. More complete collections appeared later, notably those of John Payne Collier (*Shakespeare's Library,* 1843; revised by W. Carew Hazlitt, 1875). These earlier collections have been superseded by a 7-volume version edited by Geoffrey Bullough as *Narrative and Dramatic Sources of Shakespeare* (1957–72).

Original aspects of Shakespeare's art

It has become steadily more possible to see what was original in Shakespeare's dramatic art. He achieved compression and economy by the exclusion of undramatic material. He developed characters from brief suggestions in his source (Mercutio, Touchstone, Falstaff, Pandarus), and he developed entirely new characters (the Dromio brothers, Beatrice and Benedick, Sir Toby Belch, Malvolio, Paulina, Roderigo, Lear's fool). He rearranged the plot with a view to more effective contrasts of character, climaxes, and conclusions (*Macbeth, Othello, The Winter's Tale, As You Like It*). A wider philosophical outlook was introduced (*Hamlet, Coriolanus, All's Well That Ends Well, Troilus and Cressida*). And everywhere an intensification of the dialogue and an altogether higher level of imaginative writing transformed the older work.

But, quite apart from evidence of the sources of his plays, it is not difficult to get a fair impression of Shakespeare as a reader, feeding his own imagination by a moderate acquaintance with the literary achievements of other men and of other ages. He quotes his contemporary Christopher Marlowe in *As You Like It.* He casually refers to the *Aethiopica* ("Ethiopian History") of Heliodorus (which had been translated by Thomas Underdown in 1569) in *Twelfth Night.* He read the translation of Ovid's *Metamorphoses* by Arthur Golding, which went through seven editions between 1567 and 1612. Chapman's vigorous translation of Homer's *Iliad* impressed him, though he used some of the material rather sardonically in *Troilus and Cressida.* He derived the ironical account of an ideal republic in *The Tempest* from one of Montaigne's essays. He read (in part, at least) Samuel Harsnett's *A Declaration of Egregious Popish Impostures* and remembered lively passages from it when he was writing *King Lear.* The beginning lines of one sonnet (106) indicate that he had read Edmund Spenser's poem *The Faerie Queene* or comparable romantic literature.

He was acutely aware of the varieties of poetic style that characterized the work of other authors. A brilliant little poem he composed for Prince Hamlet (5.2.115) shows how ironically he perceived the qualities of poetry in the last years of the 16th century, when poets such as John Donne were writing love poems uniting astronomical and cosmogenic imagery with skepticism and moral paradoxes. The eight-syllable lines in an archaic mode written for the 14th-century poet John Gower in *Pericles* show his reading of that poet's *Confessio amantis.* The influence of the great figure of Sir Philip Sidney, whose *Arcadia* was first printed in 1590 and was widely read for generations, is frequently felt in Shakespeare's writings. Finally, the importance of the Bible for Shakespeare's style and range of allusion is not to be underestimated. His works show a pervasive familiarity with the passages appointed to be read in church on each Sunday throughout the year, and the large number of allusions to passages in Ecclesiasticus (Wisdom of Jesus the Son of Sirach) indicates a personal interest in one of the uncanonical books.

(J.R.Br./T.Sp./Ed.)

Understanding Shakespeare

QUESTIONS OF AUTHORSHIP

Shakespeare's authorship unquestioned in his lifetime

Readers and playgoers in Shakespeare's own lifetime, and indeed until the late 18th century, never questioned Shakespeare's authorship of his plays. He was a well-known actor from Stratford who performed in London's premier acting company, among the great actors of his day. He was widely known by the leading writers of his time as well, including Ben Jonson and John Webster, both of whom praised him as a dramatist. Many other

tributes to him as a great writer appeared during his lifetime. Any theory that supposes him not to have been the writer of the plays and poems attributed to him must suppose that Shakespeare's contemporaries were universally fooled by some kind of secret arrangement.

Claims of authorship for Sir Francis Bacon

Yet suspicions on the subject gained increasing force in the mid-19th century. One Delia Bacon proposed that the author was her claimed ancestor Sir Francis Bacon, Viscount St. Albans, who was indeed a prominent writer of the Elizabethan era. What had prompted this theory? The chief considerations seem to have been that little is known about Shakespeare's life (though in fact more is known about him than about his contemporary writers), that he was from the country town of Stratford-upon-Avon, that he never attended one of the universities, and that therefore it would have been impossible for him to write knowledgeably about the great affairs of English courtly life such as we find in the plays.

The theory is suspect on a number of counts. University training in Shakespeare's day centred on theology and on Latin, Greek, and Hebrew texts of a sort that would not have greatly improved Shakespeare's knowledge of contemporary English life. By the 19th century, a university education was becoming more and more the mark of a broadly educated person, but university training in the 16th century was quite a different matter. The notion that only a university-educated person could write of life at court and among the gentry is an erroneous and indeed a snobbish assumption. Shakespeare was better off going to London as he did, seeing and writing plays, listening to how people talked. The great writers of his era (or indeed of most eras) are not usually aristocrats, who have no need to earn a living by their pens. Shakespeare's social background is essentially like that of his best contemporaries. Edmund Spenser went to Cambridge, it is true, but he came from a sail-making family. Christopher Marlowe also attended Cambridge, but his kindred were shoemakers in Canterbury. John Webster, Thomas Dekker, and Thomas Middleton came from similar backgrounds. They discovered that they were writers, able to make a living off their talent, and they (excluding the poet Spenser) flocked to the London theatres, where customers for their wares were to be found. Like them, Shakespeare was a man of the commercial theatre.

Other candidates—William Stanley, 6th earl of Derby, and Christopher Marlowe among them—have been proposed, and indeed the very fact of so many candidates makes one suspicious of the claims of any one person. The current candidate for the writing of Shakespeare's plays, other than Shakespeare himself, is Edward de Vere, 17th earl of Oxford. Oxford did indeed write verse, as did other gentlemen; sonneteering was a mark of gentlemanly distinction. Oxford was also a wretched man who abused his wife and drove his father-in-law to distraction. Most seriously damaging to Oxford's candidacy is the fact that he died in 1604. The chronology presented here, summarizing perhaps 200 years of assiduous scholarship, establishes a professional career for Shakespeare as dramatist that extends from about 1589 to 1614. Many of his greatest plays—*King Lear, Antony and Cleopatra,* and *The Tempest* to name but three—were written after 1604. To suppose that the dating of the canon is totally out of whack and that all the plays and poems were written before 1604 is a desperate argument. Some individual dates are uncertain, but the overall pattern is coherent. The growth in poetic and dramatic styles, the development of themes and subjects, along with objective evidence, all support a chronology that extends to about 1614. To suppose alternatively that Oxford wrote the plays and poems before 1604 and then put them away in a drawer, to be brought out after his death and updated to make them appear timely, is to invent an answer to a nonexistent problem.

Earl of Oxford as possible author of plays

When all is said, the sensible question one must ask is, why would Oxford want to write the plays and poems and then not claim them for himself? The answer given is that he was an aristocrat and that writing for the theatre was not elegant; hence, he needed a front man, an alias. Shakespeare, the actor, was a suitable choice. But is it plausible that a cover-up like this could have succeeded?

Shakespeare's contemporaries, after all, wrote of him unequivocally as the author of the plays. Ben Jonson, who knew him well, contributed verses to the First Folio of 1623, where (as elsewhere) he criticizes and praises Shakespeare as the author. John Heminge and Henry Condell, fellow actors and theatre owners with Shakespeare, signed the dedication and a foreword to the First Folio and described their methods as editors. In an age that loved gossip and mystery as much as any, it seems hardly conceivable that Jonson's and Shakespeare's theatrical associates shared the secret of a gigantic literary hoax without a single leak or that they could have been imposed upon without suspicion. Unsupported assertions that the author of the plays was a man of great learning and that Shakespeare of Stratford was an illiterate rustic no longer carry weight, and only when a believer in Bacon or Oxford or Marlowe produces sound evidence will scholars pay close attention.

LINGUISTIC, HISTORICAL, TEXTUAL, AND EDITORIAL PROBLEMS

Changed meaning of words since Shakespeare's time

Since the days of Shakespeare, the English language has changed, and so have audiences, theatres, actors, and customary patterns of thought and feeling.

Problems are most obvious in single words. In the 21st century *presently,* for instance, does not mean "immediately," as it usually did for Shakespeare, or *will* mean "lust," or *rage* mean "folly," or *silly* denote "innocence" and "purity." In Shakespeare's day, words sounded different, too, so that *ably* could rhyme with *eye* or *tomb* with *dumb.* Syntax was often different, and, far more difficult to define, so was response to metre and phrase. What sounds formal and stiff to a modern hearer might have sounded fresh and gay to an Elizabethan.

Ideas have changed, too, most obviously political ones. Shakespeare's contemporaries almost unanimously believed in authoritarian monarchy and recognized divine intervention in history. Most of them would have agreed that a man should be burned for ultimate religious heresies. It is the office of linguistic and historical scholarship to aid the understanding of the multitude of factors that have significantly affected the impressions made by Shakespeare's plays.

None of Shakespeare's plays has survived in his handwritten manuscript, and, in the printed texts of some plays, notably *King Lear* and *Richard III,* there are passages that are manifestly corrupt. Even if the printer received a good manuscript, small errors could still have been introduced. Compositors were less than perfect; they often "regularized" the readings of their copy, altered punctuation in accordance with their own preferences or "house" style or because they lacked the necessary pieces of type, or made mistakes because they had to work too hurriedly. Even the correction of proof sheets in the printing house could further corrupt the text, since such correction was usually effected without reference to the author or to the manuscript copy; when both corrected and uncorrected states are still available, it is sometimes the uncorrected version that is preferable. Correctors are responsible for some errors now impossible to right.

(J.R.Br./T.Sp./D.Bev.)

LITERARY CRITICISM

During his own lifetime and shortly afterward, Shakespeare enjoyed fame and considerable critical attention. The English writer Francis Meres, in 1598, declared him to be England's greatest writer in comedy and tragedy. Writer and poet John Weever lauded "honey-tongued Shakespeare." Ben Jonson, Shakespeare's contemporary and a literary critic in his own right, granted that Shakespeare had no rival in the writing of comedy, even in the ancient classical world, and that he equaled the ancients in tragedy as well, but Jonson also faulted Shakespeare for having a mediocre command of the classical languages and for ignoring classical rules. Jonson objected when Shakespeare dramatized history extending over many

years and moved his dramatic scene around from country to country, rather than focusing on 24 hours or so in a single location. Shakespeare wrote too glibly, in Jonson's view, mixing kings and clowns, lofty verse with vulgarity, mortals with fairies.

Seventeenth century. Jonson's neoclassical perspective on Shakespeare was to govern the literary criticism of the later 17th century as well. John Dryden, in *Of Dramatick Poesie* (1668) and other essays, condemned the improbabilities of Shakespeare's late romances. Shakespeare lacked decorum, in Dryden's view, largely because he had written for an ignorant age and poorly educated audiences. He excelled in "fancy" or imagination but lagged behind in "judgment." He was a native genius, untaught, whose plays needed to be extensively rewritten to clear them of the impurities of their frequently vulgar style. And, in fact, most productions of Shakespeare on the London stage during the Restoration did just that: they rewrote Shakespeare to make him more refined.

Eighteenth century. This critical view persisted into the 18th century as well. Alexander Pope undertook to edit Shakespeare in 1725, expurgating his language and "correcting" supposedly infelicitous phrases. Samuel Johnson also edited Shakespeare's works (1765), defending his author as one who "holds up to his readers a faithful mirror of manners and of life"; but, though he pronounced Shakespeare an "ancient" (supreme praise from Johnson), he found Shakespeare's plays full of implausible plots quickly huddled together at the end, and he deplored Shakespeare's fondness for punning. Even in his defense of Shakespeare as a great English writer, Johnson lauded him in classical terms, for his universality, his ability to offer a "just representation of general nature" that could stand the test of time.

Samuel Johnson applauds Shakespeare

Romantic critics. For Romantic critics such as Samuel Taylor Coleridge in the early 19th century, Shakespeare deserved to be appreciated most of all for his creative genius and his spontaneity. For Goethe in Germany as well, Shakespeare was a bard, a mystical seer. Most of all, Shakespeare was considered supreme as a creator of character. Maurice Morgann wrote such character-based analyses as appear in his book *An Essay on the Dramatic Character of Sir John Falstaff* (1777), where Falstaff is envisaged as larger than life, a humane wit and humorist who is no coward or liar in fact but a player of inspired games. Romantic critics, including Charles Lamb, Thomas De Quincey, and William Hazlitt, extolled Shakespeare as a genius able to create an imaginative world of his own, even if Hazlitt was disturbed by what he took to be Shakespeare's political conservatism. In the theatre of the Romantic era, Shakespeare fared less well, but as an author he was much touted and even venerated. In 1769 the famous actor David Garrick had instituted a Shakespeare Jubilee at Stratford-upon-Avon to celebrate Shakespeare's birthday. Shakespeare had become England's national poet.

Importance of scholarship. The late 19th and early 20th centuries saw major increases in the systematic and scholarly exploration of Shakespeare's life and works. Philological research established a more reliable chronology of the work than had been hitherto available. Edward Dowden, in his *Shakspere: A Critical Study of His Mind and Art* (1875), analyzed the shape of Shakespeare's career in a way that had not been possible earlier. A.C. Bradley's magisterial *Shakespearean Tragedy* (1904), a book that remains highly readable, showed how the achievements of scholarship could be applied to a humane and moving interpretation of Shakespeare's greatest work. As in earlier studies of the 19th century, Bradley's approach focused largely on character.

Twentieth century. Increasingly in the 20th century, scholarship furthered understanding of Shakespeare's social, political, economic, and theatrical milieu. Shakespeare's sources came under new and intense scrutiny. Elmer Edgar Stoll, in *Art and Artifice in Shakespeare* (1933), stressed the ways in which the plays could be seen as constructs intimately connected with their historical environment. Playacting depends on conventions, which must be understood in their historical context. Costuming signals meaning to the audience; so do the theatrical building, the props, the actors' gestures.

Accordingly, historical critics sought to know more about the history of London's theatres (as in John Cranford Adams's well-known model of the Globe playhouse, or in C. Walter Hodges's *The Globe Restored,* 1953), about audiences (Alfred Harbage, *As They Liked It* [1947], and Ann Jennalie Cook, *The Privileged Playgoers of Shakespeare's London, 1576–1642* [1981]), about staging methods (Bernard Beckerman, *Shakespeare at the Globe 1599–1609* [1962]), and much more. Other scholarly studies examined censorship, the religious controversies of the Elizabethan era and how those controversies affected playwriting, and the heritage of native medieval English drama. Studies in the history of ideas have examined Elizabethan cosmology, astrology, philosophical ideas such as the Great Chain of Being, physiological theories about the four bodily humours, the political theories of Machiavelli and others, the skepticism of Montaigne, and much more.

As valuable as it is, historical criticism has not been without its opponents. A major critical movement of the 1930s and 1940s was the so-called New Criticism of F.R. Leavis, L.C. Knights, Derek Traversi, Robert Heilman, and many others, urging a more formalist approach to the poetry. "Close reading" became the mantra of this movement. At its most extreme, it urged the ignoring of historical background in favour of an intense and personal engagement with Shakespeare's language: tone, speaker, image patterns, verbal repetitions, and rhythms. Studies of imagery, of rhetorical patterns, of wordplay, and still more gave support to the movement. At the commencement of the 21st century, close reading remained an acceptable approach to the Shakespearean text.

New interpretive approaches. Shakespeare criticism of the 20th and 21st centuries has seen an extraordinary flourishing of new schools of critical approach. Psychological and psychoanalytic critics such as Ernest Jones have explored questions of character in terms of Oedipal complexes, narcissism, and psychotic behaviour or, more simply, in terms of the conflicting need in any relationship for autonomy and dependence. Mythological and archetypal criticism, especially in the influential work of Northrop Frye, has examined myths of vegetation having to do with the death and rebirth of nature as a basis for great cycles in the creative process. Christian interpretation seeks to find in Shakespeare's plays a series of deep analogies to the Christian story of sacrifice and redemption.

Critical approaches to the plays

Conversely, some criticism has pursued a vigorously iconoclastic line of interpretation. Jan Kott, writing in the disillusioning aftermath of World War II and from an eastern European perspective, reshaped Shakespeare as a dramatist of the absurd, skeptical, ridiculing, and anti-authoritarian. Kott's deeply ironic view of the political process impressed filmmakers and theatre directors such as Peter Brook (*King Lear, A Midsummer Night's Dream*). He also caught the imagination of many academic critics who were chafing at a modern political world increasingly caught up in image making and the various other manipulations of the powerful new media of television and electronic communication. A number of the so-called New Historicists (among them Stephen Greenblatt, Stephen Orgel, and Richard Helgerson) read avidly in cultural anthropology, learning from Clifford Geertz and others how to analyze literary production as a part of a cultural exchange through which a society fashions itself by means of its political ceremonials. Stephen Greenblatt's *Renaissance Self-Fashioning* (1980) provided an energizing model for the ways in which literary criticism could analyze the process. Mikhail Bakhtin was another dominant influence. In Britain the movement came to be known as Cultural Materialism; it was a first cousin to American New Historicism, though often with a more class-conscious and Marxist ideology. The chief proponents of this movement with regard to Shakespeare criticism are Jonathan Dollimore, Alan Sinfield, John Drakakis, and Terry Eagleton.

Feminist criticism and gender studies. These made significant gains after 1980. Feminists, like New Historicists, were interested in contextualizing Shakespeare's writings

rather than subjecting them to ahistorical formalist analysis. Turning to anthropologists such as Claude Lévi-Strauss, feminist critics illuminated the extent to which Shakespeare inhabited a patriarchal world dominated by men and fathers, in which women were essentially the means of exchange in power relationships among those men. Feminist criticism is deeply interested in marriage and courtship customs, gender relations, and family structures. In *The Tempest,* for example, feminist interest tends to centre on Prospero's dominating role as father and on the way in which Ferdinand and Miranda become engaged and in effect married when they pledge their love to one another in the presence of a witness—Miranda's father. Plays and poems dealing with domestic strife (such as Shakespeare's *The Rape of Lucrece*) take on a new centrality in this criticism. Diaries, marriage-counseling manuals, and other such documents become important to feminist study. Revealing patterns emerge in Shakespeare's plays as to male insecurities about women, men's need to dominate and possess women, their fears of growing old, and the like. *Much Ado About Nothing* can be seen as about men's fears of being cuckolded; *Othello* treats the same male weakness with deeply tragic consequences. The tragedy in *Romeo and Juliet* depends in part on Romeo's sensitivity to peer pressure that seemingly obliges him to kill Tybalt and thus choose macho male loyalties over the more gentle and forgiving model of behaviour he has learned from Juliet. These are only a few examples. Feminist critics include, among many others, Lynda Boose, Lisa Jardine, Gail Paster, Jean Howard, Karen Newman, Carol Neely, Peter Erickson, and Madelon Sprengnether.

Feminist readings of the plays

Gender studies such as those of Bruce R. Smith and Valerie Traub also dealt importantly with issues of gender as a social construction and with changing social attitudes toward "deviant" sexual behaviour: cross-dressing, same-sex relationships, and bisexuality.

Deconstruction. The critical movement generally known as deconstruction centred on the instability and protean ambiguity of language. It owed its origins in part to the linguistic and other work of French philosophers and critics such as Ferdinand de Saussure, Michel Foucault, and Jacques Derrida. Some of the earliest practitioners and devotees of the method in the United States were Geoffrey Hartmann, J. Hillis Miller, and Paul de Man, all of Yale University. Deconstruction stressed the extent to which "meaning" and "authorial intention" are virtually impossible to fix precisely. Translation and paraphrase are exercises in approximation at best.

Deconstruction in Shakespeare criticism

The implications of deconstruction for Shakespeare criticism have to do with language and its protean flexibility of meanings. Patricia Parker's *Shakespeare from the Margins: Language, Culture, Context* (1996), for example, offers many brilliant demonstrations of this, one of which is her study of the word *preposterous,* a word she finds throughout the plays. It means literally behind for before, back for front, second for first, end or sequel for beginning. It suggests the cart before the horse, the last first, and "arsie versie," with obscene overtones. It is thus a term for disorder in discourse, in sexual relationships, in rights of inheritance, and much more. Deconstruction as a philosophical and critical movement aroused a good deal of animosity because it questioned the fixity of meaning in language. At the same time, however, deconstruction attuned readers to verbal niceties, to layers of meaning, to nuance.

Late 20th-century and early 21st-century scholars were often revolutionary in their criticism of Shakespeare. To readers the result frequently appeared overly postmodern and trendy, presenting Shakespeare as a contemporary at the expense of more traditional values of tragic intensity, comic delight, and pure insight into the human condition. No doubt some of this criticism, as well as some older criticism, was too obscure and ideologically driven. Yet deconstructionists and feminists, for example, at their best portray a Shakespeare of enduring greatness. His durability is demonstrable in the very fact that so much modern criticism, despite its mistrust of canonical texts written by "dead white European males," turns to Shakespeare again and again. He is dead, white, European, and male, and yet he appeals irresistibly to readers and theatre audiences all over the world. To many feminist critics he is as important an author as Virginia Woolf and George Eliot.

(D.Bev.)

Shakespeare on film

While Shakespeare criticism was entering a serious scholarly phase, at the end of the 19th and the start of the 20th centuries, the nature of public entertainment itself was changing dramatically. A new medium, that of moving pictures, was entering an experimental phase. Pioneer French filmmakers began to produce primitive *actualités* (*i.e.,* brief footage of parading soldiers and umbrella dancers), which were screened between the live acts in London and New York vaudeville houses.

Among these early films is a remarkable production of 1899 (still available) by the London studio of the British Mutoscope and Biograph Company: a scene from Shakespeare's *King John*—then on the boards at Her Majesty's Theatre and featuring Sir Herbert Beerbohm Tree—recorded on 68-mm film. Of four excerpts shot and later exhibited at London's Palace Theatre to promote the stage production, only the death scene (5.2), long thought lost, resurfaced in 1990 in an Amsterdam film archive. Like all silent films, the scene from *King John* would have been accompanied by some variation of live music, sound effects, phonograph records, intertitles, recitations, or supplementary lectures.

Cineastes in France, North America, Italy, and Germany soon began making other Shakespeare movies. In 1900 Sarah Bernhardt appeared on-screen at the Paris Exposition in the duel scene from *Hamlet,* and in 1907 Georges Méliès attempted to make a coherent one-reel *Hamlet* that would tell the full story. Emulating the high culture of the Comédie-Française, French filmmakers organized a Film d'Art movement that cast high-profile actors in adaptations of famous plays, which proved a somewhat limiting deference to theatre.

By 1913, however, in one of the last Film d'Art releases, *Shylock* (a version of *The Merchant of Venice*), the actors successfully adapted their stage talents to film. In Italy Giovanni Pastrone (whose monumental *Cabiria* [1914] later inspired D.W. Griffith's *Intolerance* [1916]) brought the sense of grand-opera spectacle to his *Julius Caesar* (1909).

Meanwhile, in Brooklyn, New York, the Vitagraph production company had moved the camera off the stage and into the city parks. Brooklyn's Prospect Park served as one location for *A Midsummer Night's Dream* (1909), and Central Park's Bethesda Fountain for a Veronese street in *Romeo and Juliet* (1908).

Shakespeare in New York

One of the earliest surviving feature-length movies in North America is a Shakespeare movie, M.B. Dudley's *Richard III* (1912), also rediscovered in the late 20th century. The veteran Shakespeare actor Frederick B. Warde, who played the film's Richard, toured with the movie, providing appropriate recitations and commentary.

Many film directors had difficulty moving beyond filmed stage performances. F.R. Benson's *Richard III* (1911), filmed at the Stratford Theatre, even revealed the front line of the floorboards. Other directors were more creative; E. Hay Plumb, for example, took the cast of the London Drury Lane Company to the Dorset coast to film the castle scenes in a *Hamlet* (1913) starring 60-year-old Sir Johnston Forbes-Robertson as the gloomy prince. Directors Svend Gade and Heinz Schall came up with a gender-bending *Hamlet* (1920), which starred the famous actress Asta Nielsen as a cross-dressed prince. The internationally known actor Emil Jannings played the title role in *Othello* (1922) to Werner Krauss's Iago. Krauss also portrayed Shylock in a free adaptation of *The Merchant of Venice* (1923).

Back in Hollywood, Mary Pickford played a saucy Kate in *The Taming of the Shrew* (1929), the first feature-length sound movie of Shakespeare. With her sly wink to Bianca during the "submission" speech to Petruchio, she showed how film could subvert the Shakespearean text. Warner

Brothers' *A Midsummer Night's Dream* (1935), directed by émigrés Max Reinhardt and William Dieterle, showed the influence of Weimar expressionism, but it featured the incidental music of Felix Mendelssohn and contract actors James Cagney and Mickey Rooney, who played Bottom and Puck, respectively. Almost immediately thereafter, producer Irving Thalberg and director George Cukor offered a reverential *Romeo and Juliet* (1936), with Norma Shearer and Leslie Howard and a supporting cast of actors from the Hollywood expatriate British colony. Joseph Mankiewicz and John Houseman produced a spectacular "newsreel" style *Julius Caesar* (1953) that may have been a covert attack on McCarthyism.

In Laurence Olivier's landmark *Henry V* (1944) the camera participated in the action, rather than merely recording it. Olivier began with the gritty "actualities" of an opening scene at the boisterous Globe playhouse, moving from there to a realistic 19th-century stage set for the Boar's Head Inn and then soaring off into a mythical France as portrayed in the 1490 manuscript *Les Très Riches Heures du duc de Berry.* In *Hamlet* (1948) Olivier used a probing, interrogating camera and deep-focus photography to ferret out every nook and cranny of Elsinore. His brilliant performance as Richard, duke of Gloucester, in a filmed and subsequently televised *Richard III* (1955) identified him to millions of viewers as "that bottled spider . . . this poisonous hunch-back'd toad" (1.3.241*ff.*).

In the United States, Orson Welles rivaled Laurence Olivier in the production of Shakespeare films. Despite its crudities, Welles's *Macbeth* (1948) captures the essence of the play's wild imaginings. In *Chimes at Midnight* (1966), based on the Henriad, Falstaff becomes self-referentially Welles himself, a misunderstood genius. Welles's cinematic masterpiece was *Othello* (1952; restored 1992). Its skewed camera angles and film noir texture mirror Othello's agony.

Golden Age of Shakespeare movies

In the late 1960s a Golden Age for Shakespeare movies emerged, beginning with Franco Zeffirelli's exuberant *Taming of the Shrew* (1966), featuring Richard Burton and Elizabeth Taylor. Soon thereafter Zeffirelli offered a hugely popular *Romeo and Juliet* (1968) that reinvented the young lovers (played for once by actors of an age appropriate to their roles) as alienated youth in rebellion against intransigent parents.

Roman Polanski's *Macbeth* (1971) displayed raw filmic energy and bravura. The voracious eye of Polanski's camera roams over the barnyard details of a 10th-century Scottish castle that in its squalor mirrors the inner psyches of the Macbeths. The Japanese director Kurosawa Akira presented his own version of *Macbeth* in *Throne of Blood* (1957), a translation of the play into stylized nō drama. During the same period, the Russian director Grigory Kozintsev directed a *Hamlet* (1964) and a *King Lear* (1970) that employed grim charcoal textures. The British director Peter Brook filmed his bleak *King Lear* (1970) in frozen Jutland, with Paul Scofield as the aged king. Two loose adaptations in French, *Les Amants de Vérone* (1949) and Claude Chabrol's *Ophelia* (1962), captured essences of *Romeo and Juliet* and *Hamlet.*

In the 1970s and '80s young British artists angered by "the Establishment" made transgressive Shakespeare movies. Derek Jarman's *The Tempest* (1979) filtered the play through the lens of a camp-gay sensibility that, in depicting Prospero's impossible struggle to govern benevolently in a malevolent world, shared the attitudes of Polish critic Jan Kott's influential book *Shakespeare Our Contemporary* (1966). Jarman's *Tempest* was outdone by the avant-garde antics of Celestino Coronado's *A Midsummer Night's Dream* (1984). At the same time, in other circles, orthodoxy prevailed in Peter Snell's waxworks *Julius Caesar* (1970), with Charlton Heston as Antony. Two years later Heston's own ambitious *Antony and Cleopatra* (1972) proved a better "toga epic."

An unprecedented number of expensively produced Shakespeare movies were released in the late 1980s and the '90s. A young British actor, Kenneth Branagh, rapidly assumed the mantle left by Olivier with *Henry V* (1989) and *Much Ado About Nothing* (1993). In contrast to Olivier's phlegmatic warrior figure, Branagh created a Prince Hal who was Hamlet-like in his introspection. His *Much Ado,* featuring such popular American actors as Denzel Washington and Michael Keaton, privileged the play's sentimental over its ironic side. Branagh's four-hour "uncut" *Hamlet* (1996) combined the 1623 First Folio version with passages from the 1605 quarto. With the aid of spectacular photography, Branagh used flashbacks and fades, as he did in *Henry V,* to "explain" what is left unexplained in Shakespeare's play, showing a torrid affair between Ophelia and Hamlet. The hall of mirrors in the grand palace (filmed at Blenheim Palace in Oxfordshire) underscores the tension between the worlds of illusion and reality at the heart of the play: "Seems, madam? Nay, it is. I know not 'seems,'" says Hamlet to his mother. A later offering is Branagh's amusing musical comedy version of *Love's Labour's Lost* (2000).

Franco Zeffirelli's *Hamlet*

Franco Zeffirelli, too, returned after decades to filming Shakespeare but for *Hamlet* (1990) abandoned his Italianate settings in favour of medieval English castles. In it Mel Gibson proved an action-oriented prince. The following year Peter Greenaway's beautiful but obscure *Prospero's Books* (1991), starring an octogenarian John Gielgud, pioneered not only in bringing computer-based imagery into the Shakespeare movie but also in establishing ideological and artistic independence from the classic Hollywood film.

Oliver Parker's *Othello* (1995) paired a black actor, Laurence Fishburne, as a dynamic Othello, with Irène Jacob as a plucky Desdemona, but the film as a whole—despite Branagh's menacing Iago—was disappointingly stagy. Richard Loncraine's *Richard III* (1995) presented Ian McKellen as the evil Richard in a 1930s London teetering on the edge of fascism.

The line between "high" and "low" culture became increasingly blurred with director Baz Luhrmann's postmodern *Romeo & Juliet* (1996), starring Leonardo DiCaprio and Claire Danes. The young lovers inhabit a world of drugs, cars, MTV, and violence. The high mimetic language of the play belies the ironic mise-en-scène. This melding of "high" and "low" continued not so much in the full-scale adaptations of Shakespeare as in the many derivative movies that displaced plots or snippets or echoes from Shakespeare into surprising contexts. Gus Van Sant's *My Own Private Idaho* (1991) updated the Henriad's court-tavern dualities by locating the film in Portland, Oregon, where the mayor's prodigal son falls in with dissolute street people. Al Pacino's *Looking for Richard* (1996) is a witty film essay about the history of Shakespeare's *Richard III.* An earlier Branagh film, *In the Bleak Midwinter* (1995; U.S. title *A Midwinter's Tale*), explores *Hamlet* as it is rehearsed in an abandoned church by a band of struggling actors. Other derivative movies include the cerebral *Last Action Hero* (1993), which is Pirandello-like in its interplay between *Hamlet* and the film's hero (Arnold Schwarzenegger); *10 Things I Hate About You* (1999), based on *The Taming of the Shrew;* and *The King Is Alive* (2001), in which tourists stranded in a desert perform *King Lear.*

Shakespeare in Love

At the turn of the 21st century, two costume movies, *Elizabeth* (1998) and *Shakespeare in Love* (1998), presented heavily fictionalized versions of Shakespeare's life and times. *Elizabeth,* by Pakistani director Shekhar Kapur, starred Cate Blanchett as the beleaguered queen hemmed in by a chilling array of schemers and plotters. John Madden's *Shakespeare in Love* proved the more popular movie. Its witty screenplay by Marc Norman and Tom Stoppard portrays Will Shakespeare (Joseph Fiennes) as a starving young hack with a terrible case of writer's block, struggling to write an absurd play called *Romeo and Ethel, the Pirate's Daughter.* The farcical plot, however, conceals a substrata of learned in-jokes.

The early 1990s witnessed a spate of interest in Shakespeare's comedies, not generally favoured by filmmakers. Christine Edzard's *As You Like It* (1992) displayed a gritty realism. Whereas Paul Czinner's 1936 version, starring Olivier and Elisabeth Bergner, gloried in the "poetic realism" of designer Lazare Meerson, Edzard used a daring ploy in transforming Shakespeare's Forest of Arden into a hobo jungle in East London.

Trevor Nunn followed his notable television achievements—with Janet Suzman in *Antony and Cleopatra* (first broadcast in 1974) and Judi Dench and Ian McKellen in *Macbeth* (first broadcast in 1979)—with a splendid *Twelfth Night* (1996). Shot in Cornwall, it enfolds the fragile world of Illyria within the nostalgic atmosphere of a Chekhovian comedy.

Two major versions of *A Midsummer Night's Dream*, the first directed by Adrian Noble and the second by Michael Hoffman, were released in 1996 and 1999. In Noble's flawed film, the audience experiences the action through the eyes of a small boy who dreams about the play. This frame dates at least to Jane Howell's BBC televised production of *Titus Andronicus* (1985), and it persists in Julie Taymor's *Titus* (1999). Despite some sublime visual moments, the movie is unsatisfying—neither transgressive enough in its homoerotic innuendoes nor regressive enough to suit those who prefer a more innocent approach.

Michael Hoffman's version removed the play from Shakespeare's Athens to a fin-de-siècle setting in northern Italy. The film's musical score begins conventionally enough with the incidental music by Mendelssohn but yields to an anachronistic yet delightful medley of airs from Italian grand opera. Like a true New Woman of the 1890s, feisty Helena rides a bicycle, as do other characters. The effervescent music for the ballroom scene in Giuseppe Verdi's *La traviata* enlivens the townspeople's afternoon promenade in the village square. Hoffman's movie is also a lesson in art history; the film's designer, Luciana Arrighi, drew inspiration from the Pre-Raphaelites, from Gian Lorenzo Bernini's sculptures, from Etruscan relics, and from Greek mythology.

Two versions of Shakespeare's most violent play, *Titus Andronicus*, appeared in 1999 as if to affirm that apocalypse would attend the turn of the century. The first of these, directed by Christopher Dunne, was described by its marketers as "a savage epic of brutal revenge." The film is a *Götterdämmerung* of beheading, amputation, and stabbing. But Shakespeare's language has been kept meticulously intact.

The second *Titus* was the offering by the theatrical director Taymor, who had staged the play off-Broadway in 1994. She collaborated with cinematographer Luciano Tovoli and others to make brilliant Fellini-like images out of Shakespeare's lurid melodrama. In the film Taymor's haikulike montages blur the line between illusion and reality, making the savagery aesthetically bearable. Anthony Hopkins plays Titus, Jessica Lange a passionate Tamora, and Alan Cumming the decadent and utterly villainous Saturninus.

Michael Almereyda's *Hamlet* (2000), starring Ethan Hawke, replaces the Danish court with the Denmark Corporation in Manhattan. Elsinore is a nearby luxury hotel. Hawke plays a surly Prince Hamlet disgusted by his stepfather's greed and his mother's veneer of innocence. An amateur filmmaker, Hamlet lives in a world of television and cinema, delivering the "to be or not to be" soliloquy in the Action aisle of a video store. In one of several whimsical touches, while jetting to England, Hamlet discovers Claudius's orders for his execution on the hard drive of a laptop stored in the luggage bin over the sleeping Rosencrantz and Guildenstern.

Elsinore as a luxury hotel

When all is said and done, this flourishing body of work is a singular testament to Shakespeare's universality and humanity. More than 400 years have passed since he put quill to paper, yet, centuries after he first brought them to life on the small outdoor stage near the River Thames, Shakespeare's scenes, characters, and poetry continue to fuel a rich industry for film, literary, and music scholars and critics. Ultimately, of course, Shakespeare's commercial value rests on his immeasurable ability, then and now, to captivate readers, music and theatre lovers, filmmakers, and moviegoers alike.

A SELECTED FILMOGRAPHY OF SHAKESPEARE'S WORKS

The following list, organized alphabetically by play title, includes the film title, country of origin, date of production (usually the release date), running time, production company, director, and a few of the notable actors.

Antony and Cleopatra Spain/Switz./U.K.; 1972; running time: 160 minutes; production company: Transac/Izaro/Folie; director: Charlton Heston; cast: Charlton Heston (Antony), Hildegard Neil (Cleopatra), Fernando Rey (Lepidus)

As You Like It U.K.; 1936; running time: 97 minutes; production company: Inter-Allied; director: Paul Czinner; cast: Henry Ainley (Duke Senior), Felix Aylmer (Duke Frederick), Laurence Olivier (Orlando), Elisabeth Bergner (Rosalind)

——U.K.; 1992; running time: 117 minutes; production company: Sands Films; director: Christine Edzard; cast: Andrew Tiernan (Orlando/Oliver), Emma Croft (Rosalind), Cyril Cusack (Adam), James Fox (Jaques)

Hamlet France; 1900; running time: 3 minutes; production company: Maurice; director: Clément Maurice; cast: Sarah Bernhardt (Hamlet), Pierre Magnier (Laertes)

——France; 1907; running time: 10 minutes; production company: Méliès; director: Georges Méliès; cast: Georges Méliès (Hamlet)

——U.K.; 1913; running time: 54 minutes; production company: Hepworth/Gaumont; director: E. Hay. Plumb; cast: Johnston Forbes-Robertson (Hamlet)

——Germany; 1920; running time: 117 minutes; production company: Art Film; director(s): Svend Gade/Heinz Schall; cast: Asta Nielsen (Hamlet)

——U.K.; 1948; running time: 152 minutes; production company: Two Cities Films; director: Laurence Olivier; cast: Laurence Olivier (Hamlet), Jean Simmons (Ophelia), Eileen Herlie (Gertrude)

——West Germany; 1960; running time: 127 minutes; production company: Bavaria Attelier; director: Franz Peter Wirth; cast: Maximilian Schell (Hamlet)

——film title: *Ophelia;* France; 1962; running time: 105 minutes; production company: Boreal Pictures; director: Claude Chabrol; cast: André Jocelyn (Yvan/Hamlet), Juliette Mayniel (Lucie/Ophelia), Alida Valli (Claudia Lesurf/Gertrude), Claude Cerval (Adrien Lesurf/Claudius)

——film title: *Gamlet;* U.S.S.R.; 1964; running time: 148 minutes; production company: Lenfilm; director: Grigory Kozintsev; cast: Innokenti Smoktunovsky (Hamlet)

——U.K.; 1969; running time: 117 minutes; production company: Woodfall Film Productions; director: Tony Richardson; cast: Nicol Williamson (Hamlet), Marianne Faithfull (Ophelia), Judy Parfitt (Gertrude), Anthony Hopkins (Claudius)

——U.S.; 1990; running time: 135 minutes; production company: Icon Productions; director: Franco Zeffirelli; cast: Mel Gibson (Hamlet), Helena Bonham Carter (Ophelia), Glenn Close (Gertrude), Alan Bates (Claudius)

——film title: *Last Action Hero;* U.S.; 1993; running time: 130 minutes; production company: Columbia Pictures Corporation/Oak Productions; director: John McTiernan; cast: Arnold Schwarzenegger (Jack Slater/Himself), Ian McKellen (Death), Joan Plowright (Teacher)

——film title: *A Midwinter's Tale* (*In the Bleak Midwinter*); U.K.; 1995; running time: 98 minutes; production company: Midwinter Films; director: Kenneth Branagh; cast: Richard Briers (Henry Wakefield), Joan Collins (Margaretta D'Arcy)

——U.K./U.S.; 1996; running time: 242 minutes; production company: Castle Rock Entertainment; director: Kenneth Branagh; cast: Kenneth Branagh (Hamlet), Kate Winslet (Ophelia), Julie Christie (Gertrude), Charlton Heston (Player King)

——U.S.; 2000; running time: 123 minutes; production company: Double A Films; director: Michael Almereyda; cast: Ethan Hawke (Hamlet), Diane Venora (Gertrude), Julia Stiles (Ophelia), Sam Shepard (Ghost), Bill Murray (Polonius)

Henry IV, Part 1 (also covers *Henry IV, Part 2* and *Henry V*) film title: *Chimes at Midnight;* Spain/Switz.; 1966; running time: 119 minutes; production company: Internacional Films/Alpine; director: Orson Welles; cast: Orson Welles (Falstaff), Keith Baxter (Prince Hal), John Gielgud (Henry IV), Margaret Rutherford (Mistress Quickly)

Henry IV, Part 2 (see above *Henry IV, Part 1*)

Henry V (see above *Henry IV, Part 1*)

——U.K.; 1944; running time: 137 minutes; production company: Two Cities Films; director: Laurence Olivier; cast: Lau-

rence Olivier (Henry V), Robert Newton (Pistol), Leslie Banks (Chorus), Renée Asherson (Katherine)

——U.K.; 1989; running time: 138 minutes; production company: Samuel Goldwyn/Renaissance Films; director: Kenneth Branagh; cast: Kenneth Branagh (Henry V), Derek Jacobi (Chorus), Ian Holm (Fluellen), Judi Dench (Mistress Quickly)

——film title: *My Own Private Idaho;* U.S.; 1991; running time: 102 minutes; production company: New Line Cinema; director: Gus Van Sant; cast: River Phoenix (Mike Waters), Keanu Reeves (Scott Favor), William Richert (Bob Pigeon)

Julius Caesar U.S.; 1950; running time: 90 minutes; production company: Avon Productions; director: David Bradley; cast: Charlton Heston (Mark Antony)

——U.S.; 1953; running time: 121 minutes; production company: MGM; director: Joseph L. Mankiewicz; cast: Marlon Brando (Mark Antony), James Mason (Brutus), John Gielgud (Cassius), Louis Calhern (Julius Caesar)

——U.K.; 1970; running time: 117 minutes; production company: Commonwealth United; director: Stuart Burge; cast: Charlton Heston (Mark Antony), Jason Robards (Brutus), John Gielgud (Julius Caesar), Diana Rigg (Portia)

King John U.K.; 1899; running time: 2 minutes; production company: British Mutoscope and Biograph Company; director: Sir Herbert Beerbohm Tree; cast: Sir Herbert Beerbohm Tree (King John)

King Lear film title: *Karol Lear;* U.S.S.R.; 1970; running time: 140 minutes; production company: Lenfilm; director: Grigory Kozintsev; cast: Yury Yarvet (King Lear)

——U.K./Denmark; 1970; running time: 137 minutes; production company: Filmways (London)–Athene/Laterna Films (Copenhagen); director: Peter Brook; cast: Paul Scofield (Lear), Irene Worth (Goneril), Jack MacGowran (Fool), Anne-Lise Gabold (Cordelia)

——film title: *Ran,* or *Chaos;* Japan/France; 1985; running time: 160 minutes; production company: Greenwich Film/Herald Ace/Nippon Herald; director: Kurosawa Akira; cast: Nakadai Tatsuya (Lord Ichimonji Hidetora), Tazaki Jun (Ayabe Seiji), Igawa Hisashi (Kurogane Shuri)

——film title: *The King Is Alive;* Denmark/Sweden/U.S.; 2000; running time: 110 minutes; production company: Danish Broadcasting Corporation and others; director: Kristian Levring; cast: Miles Anderson (Jack), David Bradley (Henry)

Love's Labour's Lost U.K./France/U.S.; 2000; running time: 93 minutes; production company: Arts Council of England and others; director: Kenneth Branagh; cast: Kenneth Branagh (Berowne), Nathan Lane (Costard), Richard Briers (Nathaniel), Alicia Silverstone (The Princess)

Macbeth U.S.; 1948; running time: 89 minutes; production company: Republic Pictures/Mercury Production; director: Orson Welles; cast: Orson Welles (Macbeth), Jeanette Nolan (Lady Macbeth), Dan O'Herlihy (Macduff)

——film title: *Throne of Blood;* Japan; 1957; running time: 105 minutes; production company: Toho; director: Kurosawa Akira; cast: Mifune Toshiro (Washizu Taketoki/Macbeth), Yamada Isuzu (Asaji/Lady Macbeth)

——U.K.; 1971; running time: 140 minutes; production company: Playboy Productions/Caliban Films; director: Roman Polanski; cast: Jon Finch (Macbeth), Francesca Annis (Lady Macbeth)

The Merchant of Venice film title: *Il mercante di Venezia;* Italy; 1910; running time: 8 minutes; production company: Film d'Arte Italiana; director: Gerolamo Lo Savio; cast: Ermete Novelli (Shylock), Francesca Bertini (Portia)

——film title: *Shylock;* France; 1913; running time: 33 minutes; production company: Eclipse; director: Henri Desfontaines; cast: Harry Baur (Shylock), Pépa Bonafé (Portia)

——film title: *Der Kaufmann von Venedig;* Germany; 1923; running time: 64 minutes; production company: Peter Paul Felner-Film Company; director: Peter Paul Felner; cast: Werner Krauss (Shylock), Henny Porten (Portia), Max Schreck (Duke of Venice), Carl Ebert (Antonio)

A Midsummer Night's Dream U.S.; 1909; running time: 8 minutes; production company: Vitagraph Company; director: Charles Kent; cast: Maurice Costello (Lysander), Dolores Costello (Fairy), William Ranous (Nick Bottom)

——U.S.; 1935; running time: 132 minutes; production company: Warner Brothers; director: Max Reinhardt and William Dieterle; cast: Dick Powell (Lysander), Olivia de Havilland (Hermia), Mickey Rooney (Puck), James Cagney (Nick Bottom)

——Spain/U.K.; 1984; running time: 80 minutes; production company: Cabochon; director: Celestino Coronado; cast: Lindsay Kemp (Puck), Francois Testory (Changeling)

——U.K.; 1996; running time: 105 minutes; production company: Edenwood Productions; director: Adrian Noble; cast: Lindsay Duncan (Hippolyta/Titania), Alex Jennings (Theseus/Oberon), Desmond Barrit (Nick Bottom), Osheen Jones (The Boy)

Much Ado About Nothing U.K./U.S.; 1993; running time: 110 minutes; production company: Samuel Goldwyn/Renaissance Films; director: Kenneth Branagh; cast: Kenneth Branagh (Benedick), Emma Thompson (Beatrice), Michael Keaton (Dogberry), Denzel Washington (Don Pedro)

Othello Germany; 1922; running time: 93 minutes; production company: Wörner Film; director: Dimitri Buchowetzki; cast: Emil Jannings (Othello), Werner Krauss (Iago), Ica von Lenkeffy (Desdemona)

——Morocco; 1952; running time: 91 minutes; production company: Films Marceau/Mercury Productions; director: Orson Welles; cast: Orson Welles (Othello), Micheál MacLiammóir (Iago), Suzanne Cloutier (Desdemona), Robert Coote (Roderigo)

——U.S.S.R.; 1956; running time: 108 minutes; production company: Mosfilm; director: Sergey Yutkevich; cast: Sergey Bondarchuk (Othello), Andrey Popov (Iago), Irina Skobtseva (Desdemona)

——U.K.; 1965; running time: 165 minutes; production company: BHE Films; director(s): John Dexter, Stuart Burge; cast: Laurence Olivier (Othello), Frank Finlay (Iago), Maggie Smith (Desdemona)

——U.K.; 1995; running time: 124 minutes; production company: Castle Rock/Dakota Films/Imminent Films; director: Oliver Parker; cast: Laurence Fishburne (Othello), Kenneth Branagh (Iago), Irène Jacob (Desdemona)

Richard III U.K.; 1911; running time: 16 minutes; production company: Co-operative Cinematograph; director: Frank R. Benson; cast: Frank R. Benson (Richard III)

——U.S.; 1912; running time: 55 minutes; production company: Shakespeare Film Company/Richard III Film Company; director: James Keane [Keene]; cast: Frederick Warde (Richard III), James Keane [Keene] (Richmond)

——U.K.; 1955; running time: 138 minutes; production company: London Film Productions; director: Laurence Olivier; cast: Laurence Olivier (Richard III), John Gielgud (Clarence), Ralph Richardson (Buckingham), Claire Bloom (Lady Anne)

——U.S.; 1995; running time: 105 minutes; production company: Bayly/Paré; director: Richard Loncraine; cast: Ian McKellen (Richard III), Jim Broadbent (Buckingham), Kristin Scott Thomas (Lady Anne), Annette Bening (Queen Elizabeth)

——film title: *Looking for Richard;* U.S.; 1996; running time: 109 minutes; production company: 20th Century Fox/Chal Productions/Jam Productions; director: Al Pacino; cast: Al Pacino (Richard III), Aidan Quinn (Richmond), Alec Baldwin (Clarence), Winona Ryder (Lady Anne)

Romeo and Juliet U.S.; 1936; running time: 126 minutes; production company: MGM; director: George Cukor; cast: Leslie Howard (Romeo), Norma Shearer (Juliet), John Barrymore (Mercutio), Basil Rathbone (Tybalt)

——film title: *Les Amants de Vérone;* France; 1949; running time: 110 minutes; production company: Films de France; director: André Cayatte; cast: Serge Reggiani (Romeo), Anouk Aimée (Juliet)

——film title: *Giulietta e Romeo;* U.K./Italy; 1954; running time: 138 minutes; production company: Verona Productions; director: Renato Castellani; cast: Laurence Harvey (Romeo), Susan Shentall (Juliet), Flora Robson (Nurse)

——film title: *West Side Story;* U.S.; 1961; running time: 151 minutes; production company: United Artists and others; director(s): Robert Wise/Jerome Robbins; cast: Natalie Wood (Maria), Richard Beymer (Tony)

——film title: *Giulietta e Romeo;* Italy/Spain; 1964; running time: 90 minutes; production company: Imprecine/His-

pamer Film; director: Riccardo Freda; cast: Gerald Meynier (Romeo), Rosemarie Dexter (Juliet)

——Italy/U.K.; 1968; running time: 139 minutes; production company: B.H.E./Verona Productions/Dino de Laurentiis Cinematografica; director: Franco Zeffirelli; cast: Leonard Whiting (Romeo), Olivia Hussey (Juliet), Michael York (Tybalt)

——U.S.; 1996; running time: 120 minutes; production company: Bazmark; director: Baz Luhrmann; cast: Leonardo DiCaprio (Romeo), Claire Danes (Juliet), Brian Dennehy (Montague), Paul Sorvino (Capulet)

——film title: *Tromeo and Juliet;* U.S.; 1996; running time: 102 minutes; production company: Troma Films; director(s): Lloyd Kaufman/James Gunn II; cast: Jane Jensen (Juliet), Will Keenan (Tromeo Que)

The Taming of the Shrew U.S.; 1929; running time: 68 minutes; production company: Pickford Corporation; director: Sam Taylor; cast: Mary Pickford (Katharina), Douglas Fairbanks (Petruchio)

——U.S./Italy; 1966; running time: 122 minutes; production company: Royal Films International (N.Y.) F.A.I. Production; director: Franco Zeffirelli; cast: Elizabeth Taylor (Katharina), Richard Burton (Petruchio)

——film title: *10 Things I Hate About You;* U.S.; 1999; running time: 97 minutes; production company: Jaret Entertainment and others; director: Gil Junger; cast: Heath Ledger (Patrick Verona), Julia Stiles (Katarina Stratford), Larisa Oleynik (Bianca Stratford)

The Tempest U.K.; 1979; running time: 96 minutes; production company: Boyd's Company; director: Derek Jarman; cast: Heathcote Williams (Prospero), Karl Johnson (Ariel), Toyah Willcox (Miranda)

——film title: *Prospero's Books;* U.K./Netherlands/France/Italy; 1991; running time: 124 minutes; production company: Allarts/Cine/Camera One/Penta; director: Peter Greenaway; cast: John Gielgud (Prospero), Isabelle Pasco (Miranda), Michael Clark (Caliban)

Titus Andronicus film title: *William Shakespeare's Titus Andronicus;* U.S.; 1999; running time: 147 minutes; production company: Joe Redner Film & Productions; director: Christopher Dunne; cast: Candy K. Sweet (Tamora), Lexton Raleigh (Aaron), Robert Reece (Titus)

——film title: *Titus;* U.S.; 1999; running time: 162 minutes; production company: Clear Blue Sky Productions and others; director: Julie Taymor; cast: Jessica Lange (Tamora), Anthony Hopkins (Titus Andronicus)

Twelfth Night film title: *Dvenadtsataya noch;* U.S.S.R.; 1955; running time: 90 minutes; production company: Lenfilm; director: Yakow Fried; cast: Katya Luchko (Sebastian/Viola), Anna Larionova (Olivia)

——U.K./U.S.; 1996; running time: 134 minutes; production company: Renaissance Productions; director: Trevor Nunn; cast: Imogen Stubbs (Viola), Helena Bonham Carter (Olivia), Richard E. Grant (Sir Andrew Aguecheek), Steven Mackintosh (Sebastian)

The Winter's Tale film title: *Una tragedia alla corte di Sicilia;* Italy; 1914; running time: 32 minutes; production company: Milano Films; director: Baldassare Negroni; cast: Pina Fabbri (Paulina), V. Cocchi (Leontes)

——U.K.; 1966; running time: 151 minutes; production company: Cressida/Hurst Park Productions; director: Frank Dunlop; cast: Laurence Harvey (Leontes), Jane Asher (Perdita)

(K.R.)

BIBLIOGRAPHY

Modern editions. Late 20th-century collections of Shakespeare's works include IRVING RIBNER and GEORGE LYMAN KITTREDGE (eds.), *The Complete Works of Shakespeare* (1971); SYLVAN BARNET (ed.), *The Complete Signet Classic Shakespeare* (1972); STANLEY WELLS and GARY TAYLOR (eds.), *William Shakespeare, The Complete Works* (1986, reissued as *The Complete Works,* 1998); G. BLAKEMORE EVANS and J.J. TOBIN (eds.), *The Riverside Shakespeare,* 2nd ed. (1997); DAVID BEVINGTON (ed.), *The Complete Works of Shakespeare,* 4th ed., updated (1997); and STEPHEN GREENBLATT (ed.), *The Norton Shakespeare* (1997). Three series are in progress, with plays and poems in individual volumes: STANLEY WELLS (ed.), *The Oxford Shakespeare* (1982–); PHILIP BROCKBANK (ed.), *The New Cambridge Shakespeare* (1984–); and RICHARD PROUDFOOT, ANN THOMPSON, and DAVID SCOTT KASTAN (eds.), *The Arden Shakespeare,* 3rd series (1995–).

Shakespeare biography. The following are especially informative and up-to-date: S. SCHOENBAUM, *William Shakespeare: A Documentary Life* (1975), and *William Shakespeare: Records and Images* (1981); RICHARD DUTTON, *William Shakespeare: A Literary Life* (1989); DENNIS KAY, *Shakespeare: His Life, Work, and Era* (1992); STANLEY WELLS, *Shakespeare: A Life in Drama* (1995, reissued 1997); and PARK HONAN, *Shakespeare: A Life* (1998).

Shakespearean staging, acting companies, censorship. W.W. GREG (ed.), *Dramatic Documents from the Elizabethan Playhouses: Stage Plots, Actors' Parts, Prompt Books,* 2 vol. (1931, reissued 1969); ALFRED HARBAGE, *Shakespeare's Audience* (1941, reissued 1969), and *As They Liked It* (1947, reissued 1972); G.E. BENTLEY, *The Jacobean and Caroline Stage,* 7 vol. (1941–68), and *The Professions of Dramatist and Player in Shakespeare's Time, 1590–1642* (1986); C. WALTER HODGES, *The Globe Restored* (1953, reissued 1989); PHILIP HENSLOWE, *Henslowe's Diary,* ed. by R.A. FOAKES and R.T. RICKERT (1961, reprinted 1968); M.C. BRADBROOK, *The Rise of the Common Player: A Study of Actor and Society in Shakespeare's England* (1962, reissued 1979); BERNARD BECKERMAN, *Shakespeare at the Globe, 1599–1609* (1962); ALAN C. DESSEN, *Elizabethan Drama and the Viewer's Eye* (1977), and *Recovering Shakespeare's Theatrical Vocabulary* (1995); ANN JENNALIE COOK, *The Privileged Playgoers of Shakespeare's London, 1576–1642* (1981); R.A. FOAKES, *Illustrations of the English Stage, 1580–1642* (1985); RICHARD DUTTON, *Mastering the Revels: The Regulation and Censorship of English Renaissance Drama* (1991); and ANDREW GURR, *The Shakespearean Stage, 1576–1642,* 3rd ed. (1992), and *Playgoing in Shakespeare's London,* 2nd ed. (1996).

Critical studies. *To the 1940s:* JOHN DRYDEN, *Of Dramatick Poesie* (1668); MAURICE MORGANN, *An Essay on the Dramatic Character of Sir John Falstaff* (1777); EDWARD DOWDEN, *Shakspere: A Critical Study of His Mind and Art* (1875); ELMER EDGARD STOLL, *Art and Artifice in Shakespeare* (1933); CAROLINE SPURGEON, *Shakespeare's Imagery and What It Tells Us* (1935); G. WILSON KNIGHT, *The Wheel of Fire,* 4th rev. and enlarged ed. (1949, reissued 1998), *The Imperial Theme,* 3rd ed. (1951, reprinted 1989), and *The Shakespearian Tempest,* 3rd ed. (1953, reissued 1971); SAMUEL JOHNSON, *Johnson on Shakespeare,* ed. by ARTHUR SHERBO (1968); SAMUEL TAYLOR COLERIDGE, *Coleridge on Shakespeare,* ed. by R.A. FOAKES (1971); and A.C. BRADLEY, *Shakespearean Tragedy,* 3rd ed. (1992).

From the 1940s to the 1970s: MIRIAM JOSEPH, *Shakespeare's Use of the Arts of Language* (1947, reissued 1966); ROBERT B. HEILMAN, *This Great Stage: Image and Structure in King Lear* (1948, reissued 1976); ERNEST JONES, *Hamlet and Oedipus* (1949, reissued 1976); ALFRED HARBAGE, *Shakespeare and the Rival Traditions* (1952, reissued 1970); F.R. LEAVIS, *The Common Pursuit* (1952, reissued 1984); J.A.K. THOMSON, *Shakespeare and the Classics* (1952, reissued 1978); M.M. MAHOUD, *Shakespeare's Wordplay* (1957); BERNARD SPIVACK, *Shakespeare and the Allegory of Evil* (1958); C.L. BARBER, *Shakespeare's Festive Comedy* (1959, reissued 1990); L.C. KNIGHTS, *Some Shakespearean Themes* (1959); ANNE RIGHTER, *Shakespeare and the Idea of the Play* (1962); ROBERT G. HUNTER, *Shakespeare and the Comedy of Forgiveness* (1965); MAYNARD MACK, *King Lear in Our Time* (1965); NORTHROP FRYE, *A Natural Perspective: The Development of Shakespearean Comedy and Romance* (1965, reissued 1991), and *Fools of Time: Studies in Shakespearean Tragedy* (1967, reissued 1991); NORMAN HOLLAND, *Psychoanalysis and Shakespeare* (1966); TERENCE EAGLETON, *Shakespeare and Society* (1967); JAN KOTT, *Shakespeare Our Contemporary,* 2nd ed. (1967, reprinted 1988; originally published in Polish, 1961); PHILIP EDWARDS, *Shakespeare and the Confines of Art* (1968, reprinted 1981); DEREK TRAVERSI, *An Approach to Shakespeare,* 3rd ed., rev. and expanded, 2 vol. (1968–69); ALEXANDER LEGGATT, *Shakespeare's Comedy of Love* (1974, reprinted 1990); LEO SALINGAR, *Shakespeare and the Traditions of Comedy* (1974); W. GORDON ZEEVELD, *The Temper of Shakespeare's Thought* (1974); STEPHEN ORGEL, *The Illusion of Power: Political Theater in the English Renaissance* (1975, reissued 1991); ROBERT WEIMANN, *Shakespeare and the Popular Tradition in the Theater* (1978, reissued 1987; originally published in German, 1967); and JULIET DUSINBERRE, *Shakespeare and the Nature of Women,* 2nd ed. (1996).

From 1980 to 2000: STEPHEN GREENBLATT, *Renaissance Self-Fashioning* (1980), and *Hamlet in Purgatory* (2001); COPPELIA KAHN, *Man's Estate: Masculine Identity in Shakespeare* (1981); RICHARD P. WHEELER, *Shakespeare's Development and the Problem Comedies: Turn and Counter-Turn* (1981); ARTHUR KIRSCH, *Shakespeare and the Experience of Love* (1981), and *The Passions of Shakespeare's Tragic Heroes* (1990); DAVID SCOTT KASTAN, *Shakespeare and the Shapes of Time* (1982); NORTHROP FRYE, *The Myth of Deliverance: Reflections on Shakespeare's*

Problem Comedies (1983, reissued 1993); LISA JARDINE, *Still Harping on Daughters: Women and Drama in the Age of Shakespeare* (1983); MARIANNE NOVY, *Love's Argument: Gender Relations in Shakespeare* (1984); ROBERT N. WATSON, *Shakespeare and the Hazards of Ambition* (1984); JOHN DRAKAKIS (ed.), *Alternative Shakespeares* (1985); PETER ERICKSON, *Patriarchal Structures in Shakespeare's Drama* (1985); CAROL THOMAS NEELY, *Broken Nuptials in Shakespeare's Plays* (1985); W. THOMAS MacCARY, *Friends and Lovers: The Phenomenology of Desire in Shakespearean Comedy* (1985); TERENCE EAGLETON, *William Shakespeare* (1986); JOEL FINEMAN, *Shakespeare's Perjured Eye: The Invention of Poetic Subjectivity in the Sonnets* (1986); STANLEY CAVELL, *Disowning Knowledge in Six Plays of Shakespeare* (1987); JEAN E. HOWARD and MARION F. O'CONNOR (eds.), *Shakespeare Reproduced: The Text in History and Ideology* (1987); LEAH MARCUS, *Puzzling Shakespeare: Local Reading and Its Discontents* (1988); STEVEN MULLANEY, *The Place of the Stage: License, Play, and Power in Renaissance England* (1988); ANIA LOOMBA, *Gender, Race, Renaissance Drama* (1989); JONATHAN DOLLIMORE, *Radical Tragedy: Religion, Ideology, and Power in the Drama of Shakespeare and His Contemporaries*, 2nd ed. (1989); ANNABEL PATTERSON, *Shakespeare and the Popular Voice* (1989); KAREN NEWMAN, *Fashioning Femininity and English Renaissance Drama* (1991); BRUCE R. SMITH, *Homosexual Desire in Shakespeare's England* (1991); JANET ADELMAN, *Suffocating Mothers: Fantasies of Maternal Origin in Shakespeare's Plays* (1992); VALERIE TRAUB, *Desire and Anxiety: Circulations of Sexuality in Shakespearean Drama* (1992); ALAN SINFIELD, *Faultlines: Cultural Materialism and the Politics of Dissident Reading* (1992); LINDA CHARNES, *Notorious Identity: Materializing the Subjective Shakespeare* (1993); BRIAN VICKERS, *Appropriating Shakespeare: Contemporary Critical Quarrels* (1993); MEREDITH SKURA, *Shakespeare the Actor and the Purposes of Playing* (1993); MAYNARD MACK, *Everybody's Shakespeare: Reflections Chiefly on the Tragedies* (1993); GAIL KERN PASTER, *The Body Embarrassed: Drama and the Disciplines of Shame in Early Modern England* (1993); LARS ENGLE, *Shakespearean Pragmatism: Market of His Time* (1993); JEAN E. HOWARD, *The Stage and Social Struggle in Early Modern England* (1994); KIM F. HALL, *Things of Darkness: Economies of Race and Gender in Early Modern England* (1995); PATRICIA PARKER, *Shakespeare from the Margins: Language, Culture, Context* (1996); and PHILIPPA BERRY, *Shakespeare's Feminine Endings: Disfiguring Death in the Tragedies* (1999). (D.Bev.)

Shakespeare on film. ROBERT HAMILTON BALL, *Shakespeare on Silent Film: A Strange Eventful History* (1968), is the definitive work on silent Shakespeare movies. Supplementing it is WILLIAM URICCHIO and ROBERTA E. PEARSON, *Reframing Culture: The Case of the Vitagraph Quality Films* (1993). KENNETH S. ROTHWELL and ANNABELLE HENKIN MELZER, *Shakespeare on Screen: An International Filmography and Videography* (1990), lists more than 700 varieties of Shakespeare movies. One of the best critical surveys is JACK J. JORGENS, *Shakespeare on Film* (1977, reprinted 1991), which should be read along with ROGER MANVELL, *Shakespeare and the Film*, rev. and updated ed. (1979). Late 20th-century studies include BERNICE W. KLIMAN, *Hamlet: Film, Television and Audio Performance* (1988); ANTHONY DAVIES, *Filming Shakespeare's Plays* (1988); JOHN COLLICK, *Shakespeare Cinema and Society* (1989); PETER S. DONALDSON, *Shakespearean Films/Shakespearean Directors* (1990); SAMUEL CROWL, *Shakespeare Observed* (1992); HERBERT R. COURSEN, *Shakespeare in Production* (1996); RICHARD BURT, *Unspeakable Shaxxxpeares* (1998); KENNETH S. ROTHWELL, *A History of Shakespeare on Screen* (1999); MICHAEL ANDEREGG, *Orson Welles, Shakespeare, and Popular Culture* (1999); ROBERT F. WILLSON, JR., *Shakespeare in Hollywood, 1929–1956* (2000); and KATHY HOWLETT, *Framing Shakespeare on Film* (2000). (K.R.)

Shanghai

Shanghai (Wade-Giles romanization Shang-hai, Pinyin Shanghai), whose name literally means "on the sea," is one of the world's largest seaports and a major industrial and commercial centre of the People's Republic of China. It is located on the coast of the East China Sea between the mouth of the Yangtze River to the north and the bays of Hangchow and Yü-p'an to the south. The municipality covers a total area of 2,383 square miles (6,185 square kilometres), which includes the city itself, surrounding suburbs, and an agricultural hinterland; it is also China's most populous urban area.

Shanghai was the first Chinese port to be opened to Western trade, and it long dominated the nation's commerce. Since the Communist victory in 1949, however, it has become an industrial giant whose products supply China's growing domestic demands. The city has also undergone extensive physical changes with the establishment of industrial suburbs and housing complexes, the improvement of public works, and the provision of parks and other recreational facilities. Shanghai has attempted to eradicate the economic and psychological legacies of its exploited past through physical and social transformation to support its major role in the modernization of China.

The article is divided into the following sections:

Physical and human geography 273
  The landscape 273
    Site
    Climate
    Layout
  The people 274
  The economy 274
    Industry
    Commerce
    Finance and trade
    Transportation
  Administrative and social conditions 276
    Government
    Public utilities
    Health
    Education
  Cultural life 276
History 277
  Evolution of the city 277
  The 20th century 277
Bibliography 277

Physical and human geography

THE LANDSCAPE

The city's composition

Site. The province-level municipality (*shih*) of Shanghai is bordered by Kiangsu Province on the north and west and Chekiang Province on the south. It includes the city of Shanghai; the nine mainland counties (*hsien*) of Pao-shan, Chia-ting, Ch'ing-p'u, Sung-chiang, Chin-shan (Chu-ching), Shang-hai (Hsin-chuang), Feng-hsien (Nan-ch'iao), Nan-hui, and Ch'uan-sha; and approximately 30 islands in the mouth of the Yangtze and offshore to the southeast in the East China Sea. The largest island, Ch'ung-ming, has an area of 270 square miles (700 square kilometres), extends more than 40 miles (65 kilometres) upstream from the mouth of the Yangtze, and is the 10th county in the greater Shanghai Municipality.

The mainland portion of the city lies on an almost level deltaic plain with an average elevation of 10 to 16 feet (three to five metres) above sea level. It is crisscrossed by an intricate network of canals and waterways that connect the municipality with the T'ai Hu region to the west.

Climate. The city's maritime location fosters a mild climate characterized by minimal seasonal contrast. The average annual temperature is about 58° F (14° C); the July maximum averages about 80° F (27° C), and the average January minimum is about 37° F (3° C). About 45 inches (1,140 millimetres) of precipitation fall annually, with the heaviest rainfall in June and the lightest in December.

Layout. As China's main industrial centre, Shanghai has serious air, water, and noise pollution. Industrial relocation and construction in the suburbs since the 1950s initially helped alleviate central city air pollution, although high population density and mixed industrial-residential land use continued to cause problems. The Wu-sung (Suchou) Chiang (River) and Huang-p'u Chiang, which flow through the city, are severely polluted from industrial discharges, domestic sewage, and ships' wastes; nonetheless, the Huang-p'u is Shanghai's main water source. Environmental protection and urban cleanliness is enhanced by industrial and solid waste resource recovery operations run by a municipal corporation. More than 1,000 different materials are recycled, including plastic, chemical fibre, and residues, machine components, oil and grease, rags, human hair, and animal bones.

The municipality radiates toward the north, west, and south from the confluence of the Wu-sung, and the Huang-p'u, a tributary of the Yangtze. Surrounding the central core is a transitional zone on both banks of the Huang-p'u, which encompasses a partially rural area of about 160 square miles. The suburban industrial complexes of Wu-sung to the north and Minhsing to the south were annexed in 1980. The banks of the Wu-sung, an important inland waterway connection to the interior hinterland, are occupied by a westward arterial extension of the transitional zone. To the south, however, the transitional zone terminates abruptly a few miles south of the central Shanghai urban core, at the Huang-p'u. P'u-tung, directly east across the Huang-p'u from the central business district, was founded in 1870 as one of the earliest industrial areas; it was also notorious as the city's most extensive and appalling slum. Several of the post-1949 industrial workers' residential complexes are now located there, and it is part of the Huang-p'u district.

Downtown Shanghai. The physical perspective of downtown Shanghai is much the same as in the pre-Communist period. Because of the policy of developing integrated residential and industrial complexes in suburban areas, central city development and renewal has been given low priority. Many of the pre-World War II buildings, which housed foreign commercial concerns and diplomatic missions, still dominate the area.

Extending southward and westward from the confluence of the Wu-sung and Huang-p'u rivers, central Shanghai has a gridded street pattern and includes the area originally contained within the British concession. The area is bounded on the east along the Huang-p'u by Chung-shan Tung Lu (Chung-shan Tung Road); on the west by Hsi-tsang Chung Lu; and on the south by Yen-an Tung Lu, which was built on the former Yang-ching-p'ang Canal that separated the British from the French concessions. Chung-shan Tung Lu has several hotels, the central administrative offices of Shanghai, and a residence for foreign seamen. On the main commercial artery, Nan-ching Tung Lu, which runs westward from the eastern road, lies Shanghai's largest retail establishment—the Shanghai Number One Department Store—as well as restaurants, hotels, and the central communications building.

The Hung-k'ou district lies to the north and east of the Wu-sung Chiang. It was originally developed by American and Japanese concessionaires and in 1863 was combined with the British concession to the south to create the International Settlement. It is an important industrial area, with shipyards and factories spread out along the bank of the Huang-p'u in the eastern section of the district.

The Bund, a wide boulevard of buildings along the Huang-p'u Chiang in Shanghai.
Peter Carmichael—Aspect Picture Library

Its best known building, the Shang-hai Ta-hsia (Shanghai Mansions Hotel), overlooks the Huang-p'u.

The old Chinese city, which is now part of central Shanghai, is characterized by a random and labyrinthine street pattern. Until the early 20th century the area was surrounded by a three-mile wall. It is now circumscribed by the two streets of Jen-min Lu and Chung-hua Lu, which follow the course of the original wall; and it is bisected by the main north–south artery, Ho-nan Nan Lu (South Ho-nan Road).

Western Shanghai is primarily residential in character and is the site of the Industrial Exhibition Hall. To the southwest, the district of Hsü-hui, formerly Ziccawei, became a centre of Christian missionary activity in China in the 17th century. During the late 1800s, Jesuit priests established a major library, a printing establishment, an orphanage, and a meteorological observatory in the area.

Land-use patterns in metropolitan Shanghai mirror pre-1949 real-estate market conditions. Much of the high-value land given over to industrial plants, warehouses, and transport facilities lies close to the Huang-p'u and Wu-sung rivers. South of the Wu-sung, which is traversed by about 20 bridges within the city, residential areas extend south from the industrial strip to the Huang-p'u. North of the Wu-sung, residential areas are less clearly demarcated, and there is a more gradual merging of city and country in the transitional zone. Continuous urban settlement is bounded on the north by the two major east–west arteries of Chung-shan Pei Lu and Ssu-ping Lu.

Retail trade is concentrated in the old central business district, although the volume of trade conducted there has diminished with the establishment of the industrial satellite towns and villages on the periphery of Shanghai.

Housing. Shanghai has made considerable progress since 1949 in providing housing for its growing population. Construction of integrated, self-sufficient residential complexes in conjunction with industrial, agricultural, and commercial development throughout metropolitan and suburban Shanghai has helped disperse population from the overcrowded central city and has led to dramatic changes in the urban and suburban landscape. Thousands of families in urban districts, however, remain inadequately housed, and shanties persist in some areas.

The concept of state-supported housing was introduced in 1951 with the development of Ts'ao-yang Hsin Ts'un (Ts'ao-yang New Village) in an existing industrial zone on Shanghai's western periphery. Following the construction of the Ts'ao-yang Hsin Ts'un, many other residential complexes have been built. Some of them were constructed with the partial support of government bureaus or industrial enterprises to satisfy the needs of their employees. Two of the earliest complexes in this category were the Railroad Village and the Post and Telegraph Village.

Five major housing developments were built in the former slum area of Yang-shu-p'u. These include the villages of An-shan, K'ung-chiang, Ch'ang-pai, Feng-ch'eng, and the Feng-nan Erh Ts'un. Other complexes are those at P'eng-p'u, Chen-ju, I-ch'uan, Jih-hui, and Chiang-wan. Some of these are in relatively remote suburban locations in the transitional and hinterland zones near older rural marketing centres. The P'eng-p'u workers' housing project is typical. Those who work in nearby factories live in a garden-apartment complex that includes apartment buildings, administrative offices, workshops, clinics, and a nursery. The adjacent fields supply wheat, clover, beans, cabbage, melons, and rapeseed (for cooking oil) for consumption by the inhabitants of the complex.

THE PEOPLE

The greater municipality can be divided into three distinct population zones—the densely populated central city, the transitional zones, and the rural hinterland, which is one of the world's most densely settled agricultural areas.

Within metropolitan Shanghai, there are few, if any, concentrations of ethnic minority groups. The majority of the population is of Han Chinese origin.

THE ECONOMY

China's leading industrial centre

Industry. Shanghai has become the nation's leading industrial and manufacturing centre because of a distinctive combination of factors. These include the availability of a large, highly skilled, and technologically innovative work force; a well-grounded and broadly based scientific-research establishment supportive of industry; a tradition of cooperation among producers; and excellent internal and external communication and supply facilities.

The iron and steel industry was one of the earliest to be established in China. In the 1950s the blast-furnace capacity of the industry was enlarged, and attempts were made to integrate the operations of the iron and steel industry more closely with the machine-manufacturing industry.

Shanghai's machine and machine tool industry is especially important in China's modernization plans. Among the varieties of industrial equipment produced are multiple-use lathes, wire-drawing dies, and manufacturing equipment for assembling computers and other electronic devices, precision instruments, and polymer synthetics.

The chemical and petrochemical industries are almost fully integrated, and there is increasing cooperation among individual plants in the production and supply of chemical raw materials for plastics, synthetic fibres, dyes, paint, pharmaceuticals, agricultural pesticides, chemical fertilizers, synthetic detergents, and refined petroleum products. Heavy industry (especially metallurgical and chemical) predominated until the late 1970s. Light industry is now favoured in an effort to reduce pollution, alleviate transport congestion, and compensate for energy and raw material shortages associated with heavy industry.

The textile industry has been reorganized to assure efficient utilization of the mills' productive capacity at all stages of the manufacturing process. The textile mills cooperate in their use of raw materials and have estab-

Central Shanghai and (inset) its metropolitan area.

lished cooperative relationships with plants that manufacture rubber shoes, tires, zippers, industrial abrasives, and conveyor belts.

Shanghai is also a primary source of a wide variety of consumer goods such as watches, cameras, radios, fountain pens, glassware, stationery products, leather goods, and hardware. Factories producing such goods have made a special effort to meet consumer demands and to produce durable and attractive products.

Commerce. The retail trade in manufactured consumer goods is managed by the First Commercial Bureau. A number of commercial corporations under the Bureau are responsible, in turn, for the wholesaling, distribution, and warehousing of specific commodity groups. A separate corporation manages the larger retail stores, while the smaller retail establishments and some specialized wholesaling organizations are controlled by local commerce bureaus in the various districts of the city.

Banking operations

Finance and trade. Shanghai's two major banks—the People's Construction Bank and the Bank of China—function as administrative organs of the Ministry of Finance. They are responsible for the disbursement and management of capital investment funds for state enterprises. Two British banks, the Hong Kong and Shanghai Banking Corporation and the Chartered Bank, along with other foreign banks, maintain Shanghai branch offices that underwrite foreign trade transactions and exchange foreign currency in connection with trading operations. Remittances from Chinese living abroad (mainly in Hong Kong and in a number of Southeast Asian countries) are

managed and collected by several overseas Chinese banks.

Industrial products are exported from Shanghai to all parts of China. Imports are mainly unprocessed food grains, petroleum and coal, construction materials, and such industrial raw materials as pig iron, salt, raw cotton, tobacco, and oils. In domestic trade, Shanghai still imports more than it exports. In foreign trade, however, the value of exported commodities exceeds that of imported goods, and the proportion of manufactured exports is steadily increasing.

Transportation. Shanghai is China's major transport centre. The central city is both a sea and river port, with the Huang-p'u Chiang serving as an excellent harbour; at high tide, oceangoing vessels can sail up the river to the city.

The port of Shanghai

In the early 1950s, the harbour was divided into a number of specialized sections. P'u-tung, on the east bank of the Huang-p'u and in the Huang-p'u district, is used for the storage of bulk commodities and for transportation maintenance and repair facilities, while P'u-hsi, in the Nan-shih District on the west bank, and Fu-hsing Tao (Fu-hsing Island) are the sites of general cargo wharves. Kao-yang Lu Wharf and I-hui contain general and bulk cargo wharves, and the Wu-sung Chiang is lined with riverine and small-craft terminals and cargo-handling facilities. Ocean terminals were constructed after 1952 at Jih-hui Chiang, south of the city, and at Chang-hua-peng, to the north at Wu-sung; a third passenger-and-freight terminal for Yangtze River and coastal traffic was opened in 1982.

Heavily used inland-waterway connections, via the Wu-sung Chiang, and an extensive canal network are maintained with Su-chou (Soochow), Wu-hsi, and Yang-chou in Kiangsu Province, and with Hang-chou (Hangchow), in Chekiang Province.

Railway facilities

The railway network reflects the efforts that have been made since 1949 to reorient the city's industrial economy to balance export and domestic development needs. Before World War II, Shanghai was the terminus of two major rail lines south of the Yangtze—the Hu-ning line, from Nanking to Shanghai, and the Hu-han-jung line, from Shanghai to the port of Ning-po in Chekiang Province. A short spur line also ran from Shanghai to Wu-sung. Additional spur lines, built since 1949, connect the industrial districts to the main trunk routes. These spurs include the Wen-tsao-pin and Min-hang lines and several short spurs emanating from Nan-hsiang (just west of the central city) to Ho-chia-wan, Peng-p'u, T'ao-p'u, and the Shanghai General Petrochemical Plant at Jinshanwei to the south.

Shanghai is served by two airports. The older Lung-hua Airport, a few miles south of the city, is used mainly for domestic flights; the Hung-ch'iao International Airport, southwest of Shanghai, is one of China's busiest. Intraurban transport by electric trolleybus, trolley, and motorbus has been substantially improved since 1949.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. As a first-order, province-level administrative unit, Shanghai Municipality is, in theory, directly controlled by the central government in Peking. It is difficult, however, to gauge the precise nature of this relationship. Since the Cultural Revolution of the late 1960s, China's administrative apparatus at all levels of the hierarchy has been in a process of readjustment so as to bring governmental organization in line with political reality. In 1967, at the beginning of the Cultural Revolution, the Shanghai Municipal Revolutionary Committee was established as the top governing body in the municipality after a chaotic period in which a number of popular-based revolutionary organizations seized control of the city for brief periods. The committee at that time was composed of representatives of the army, the mass revolutionary organizations, and some former Communist Party officials. By the mid-1970s, this was replaced by a municipal government made up of commissions, offices, and bureaus responsible to the Shanghai People's Congress, an elected body. These units serve both policy advisory and administrative functions and function as administrative links to both the national government in Peking as well as the local governing bodies.

Public utilities. Modern public works improvements include the installation and improvement of drainage and sewage treatment facilities, public water supply systems, street lights, and public refuse bins. Roads have been widened and repaired, flood walls constructed in low-lying areas subject to tidal inundation, and housing built. The sea walls surrounding Shanghai have also been strengthened and enlarged; two long sea walls extend east of the Huang-p'u for a total of more than 13 miles (21 kilometres).

Shanghai is also one of China's major electric power generating centres. Electricity can be generated by coal-fired thermal plants, and the Shanghai area is linked via a major transmission network with Nanking to the northwest and with Hangchow and Hsin-an-chiang (the site of a hydroelectric generating facility) in Chekiang Province to the southwest. China's largest gas works is located at Lung-hua. Increased energy demands for industry and domestic use in the early 1980s led to a decision by the national authorities to construct one of China's first two nuclear power plants in Shanghai.

Health. Shanghai's health-care facilities range from thousands of small clinics associated with factories, schools, retail establishments, and government offices to numerous major research and teaching hospitals. Most hospitals have facilities for practicing and teaching both traditional Chinese and Western medicine. Medical schools have concentrated on the training of "barefoot" doctors, practitioners with sufficient medical skills to supply basic care to people in rural areas.

Education. Shanghai is China's leading centre of higher education and scientific research. There are numerous universities and other institutions of higher learning—including Fu-tan, Chiao-t'ung, T'ung-chi, and the Hua-tung Shih-fan Ta-hsueh—as well as technical and higher education institutes. Many factories have affiliated work-study colleges to equip workers for more highly skilled jobs. In 1960 the Shanghai Municipal Part-Work Part-Study Industrial University was established through the cooperation of more than 1,000 industrial establishments. A large segment of the city's total work force is enrolled in one of these schools.

Work and college programs

The Shanghai Branch of the Chinese Academy of Sciences, China's leading scientific research and development body, is located in Shanghai. During the Cultural Revolution, practical applications of scientific work in agriculture and industry were encouraged. Since the late 1970s, extensive research investments have been made in such high technology areas as nuclear energy, computers, semiconductors, laser and infrared technology, and satellites.

CULTURAL LIFE

Shanghai's cultural attractions include museums, historical sites, and scenic gardens. The Shanghai Museum of Art and History houses an extensive collection of bronzes, ceramics, and other artifacts dating over several thousand years. The Shanghai Revolutionary History Memorial Hall displays photographs and objects that trace the city's evolution. The Ta Shih-chieh ("Great World"), founded in the 1920s, is Shanghai's leading theatrical centre and offers folk operas, dance performances, plays, story readings, and specialized entertainment forms typical of China's national minority groups. The city also has many workers' and children's recreational clubs and several large motion picture theatres, including the Kuang-ming Theatre.

The old Chinese city houses the 16th-century Yü-Yuan Garden (Garden of the Mandarin Yü), an outstanding example of late Ming garden architecture, and the Former Temple of Confucius. Other points of attraction are the Ch'ing dynasty Lung-hua Pagoda, the Industrial Exhibition Hall, and the tomb and former residence of Lu Hsün, a 20th-century revolutionary writer.

The major publishing houses of Shanghai are a branch of the People's Literature Publishing House (at Peking) and the People's Educational Publishing House. In addition to the large branch of the library of the Chinese Academy of Sciences, Shanghai has numerous other libraries. Shanghai's art and music schools include a branch of the Central Conservatory (Tientsin), the Shanghai Con-

servatory, and the Shanghai Institute of Drama. There are also ballet and opera companies, symphonies, and puppet troupes.

Parks and playing fields were expanded after 1949. Two of the earliest to be opened for public use were the People's Park in central Shanghai and the Huang-p'u Park on the shore of the Huang-p'u Chiang. Every section of the city has parks and playing fields. Among the largest are the Hung-k'ou Arboretum and Stadium in the north; the Peace Park and Ho-p'ing Park and playing field in the northeast; the P'u-tung Park in eastern Shanghai, the Hunan and Fu-hsing parks in the south, and the Chung-shan Park on the western periphery of the central city.

History

EVOLUTION OF THE CITY

As late as the 5th to 7th century AD the Shanghai area, then known as Shen or Hu Tu, was sparsely populated and undeveloped. Despite the steady southward progression of Chinese settlement, the exposed deltaic position of the area retarded its economic growth.

During the Sung dynasty (960–1126) Shanghai emerged from its somnolent state as a small, isolated fishing village. The area to the west around T'ai Hu (T'ai Lake) had developed a self-sustaining agricultural economy on reclaimed land and was stimulated by an increase in population resulting from the southward migration of Chinese fleeing the invading Mongols in the north. The natural advantages of Shanghai as a deepwater port and shipping centre were recognized as coastal and inland shipping expanded. By the beginning of the 11th century, a customs office was established; and by the end of the 13th century, Shanghai was designated as a county seat and placed under the jurisdiction of Kiangsu Province.

During the Ming dynasty (1368–1644), roughly 70 percent of the cultivated land around Shanghai was given to the production of cotton for the city's cotton- and silk-spinning industry. By the middle of the 18th century more than 20,000 persons worked as cotton spinners.

The impact of Western imperialism

After the 1850s, the predominantly agricultural focus of the economy was quickly transformed. At this time the city became the major Chinese base for commercial imperialism by nations of the West. Following their humiliating defeat by Great Britain in 1842, the Chinese surrendered Shanghai and signed the Treaty of Nanking, which opened the city to unrestricted foreign trade. The British, French, and Americans took possession of designated areas in the city within which they were granted special rights and privileges, and the Japanese received a concession in 1895 under the terms of the Treaty of Shimonoseki.

The opening of Shanghai to foreign business immediately led to the establishment of major European banks and multipurpose commercial houses. The city's prospects as a leading centre of foreign trade were further enhanced when Canton, a rival port in the southeastern coastal province of Kwangtung, was cut off from its hinterland by the Taiping Rebellion (1850–64). Impelled by this potential threat to the uninterrupted expansion of their commercial operations in China, the British obtained rights of navigation on the Yangtze in 1857. As the natural outlet for the vast hinterland of the Lower Yangtze, Shanghai rapidly grew to become China's leading port and by 1860 accounted for about 25 percent of the total shipping tonnage entering and departing the country.

Shanghai did not, however, show promise of becoming a major industrial centre until the 1890s. Except for the Chiang-nan Arsenal organized by the Ch'ing dynasty (1644–1911) in the early 1860s, most industrial enterprises were small-scale offshoots of the larger foreign trading houses. As the flow of foreign capital steadily increased after the Sino-Japanese War of 1894–95, light industries were established within the foreign concessions, which took advantage of Shanghai's ample and cheap labour supply, local raw materials, and inexpensive power.

THE 20TH CENTURY

By contrast, local Chinese investment in Shanghai's industry was minimal until World War I diverted foreign capital from China. From 1914 through the early 1920s, Chinese investors were able to gain a tenuous foothold in the scramble to develop the industrial economy. This initial involvement was short-lived, however, as the post-World War I resurgence of Western and Japanese economic imperialism—followed closely by the Depression of the 1930s—overwhelmed many of the newly established Chinese industries. Competition became difficult, as cheaper foreign goods were dumped on the Shanghai market, and labour was attracted to relatively higher-paying jobs in foreign-owned factories. Prior to the Sino-Japanese War of 1937–45 the Japanese had gained control over about half of the city's yarn-spinning and textile-weaving capacity.

The 1920s were also a period of growing political awareness in Shanghai. Members of the working class, students, and intellectuals became increasingly politicized as foreign domination of the city's economic and political life became ever more oppressive. When the agreements signed by the United Kingdom, the United States, and Japan at the Washington Conference of 1922 failed to satisfy Chinese demands, boycotts of foreign goods were instituted.

The Chinese Communist Party was founded in Shanghai in 1921, and four years later the Communist Party led the "May 30" uprising of students and workers. This massive political demonstration was directed against feudalism, capitalism, and official connivance in foreign imperialistic ventures. The student–worker coalition actively supported the Nationalist armies under Chiang Kai-shek, but the coalition and the Communist Party were violently suppressed by the Nationalists in 1927. (B.Bo.)

Founding of the Chinese Communist Party

Shanghai was occupied by the Japanese during the Sino-Japanese War of 1937–45, and the city's industrial plants suffered extensive war damage. The city's economy was hit with inflation until 1949, and from that time until China's economic reforms of the late 1970s Shanghai experienced little economic progress.

Since the 1990s the city's economy has been transformed. Central to this remarkable achievement was the development of the new P'u-tung economic zone across the Huang-p'u River from the city's downtown business area. As part of this development, the P'u-tung International Airport, 18 miles (29 kilometres) east of the city, was completed in October 1999. During this period the city's industrial output has expanded more than threefold, while the service sector, as a proportion of the city's total economy, grew from less than one-third to almost half.

BIBLIOGRAPHY. BETTY PEJ-T'I WEI, *Shanghai: Crucible of Modern China* (1987), a survey to 1949; and HARRIET SERGEANT, *Shanghai: Collision Point of Cultures: 1918–1939* (1990), are two popular histories. RHOADS MURPHEY, *Shanghai: Key to Modern China* (1953), is a study of Shanghai's pre-World War II political and economic organization. NICHOLAS R. CLIFFORD, *Spoilt Children of Empire: Westerners in Shanghai and the Chinese Revolution of the 1920s* (1991), looks at the foreign community in Shanghai. SAM TATE and IAN MCLACHLAN, *Shanghai: 1949: The End of an Era* (1989), is a collection of photographs of the city as it was seized by Communist troops. For the post-1949 period, NEALE HUNTER, *Shanghai Journal* (1969), recounts the author's experiences as an English teacher in the Shanghai Foreign Language Institute during the Cultural Revolution.

Economic and political developments are treated in JOSEPH FEWSMITH, *Party, State, and Local Elites in Republican China: Merchant Organizations and Politics in Shanghai, 1890–1930* (1985); CHRISTOPHER HOWE, "The Level and Structure of Employment and the Sources of Labor Supply in Shanghai, 1949–1957"; and LYNN T. WHITE III, "Shanghai's Polity in Cultural Revolution," in J.W. LEWIS (ed.), *The City in Communist China* (1971). EMILY HONIG, *Sisters and Strangers* (1986), examines the lives of women working in the Shanghai cotton mills from 1919 to 1949. CHRISTIAN HENRIOT, *Shanghai, 1927–1937: Municipal Power, Locality, and Modernization* (1993), studies the government of Shanghai during the Nanking regime. A carefully documented collection of papers on Shanghai's political life, economic development, cultural and ideological milieu, and spatial development is brought together by CHRISTOPHER HOWE (ed.), *Shanghai: Revolution and Development in an Asian Metropolis* (1981). An account of late 20th-century changes is provided in HAROLD D. FOSTER *et al.*, *The Dragon's Head: Shanghai, China's Emerging Megacity* (1998). A useful source of statistical information is THE SHANGHAI STATISTICS BUREAU, *Shanghai Statistical Yearbook* (annual). (B.Bo./Ed.)

Shintō

Shintō is the name given to indigenous religious beliefs and practices of Japan. The word Shintō literally means "the way of *kami*" (*kami* means "mystical," "superior," or "divine," generally sacred or divine power, specifically the various gods or deities); it came into use in order to distinguish indigenous Japanese beliefs from Buddhism, which had been introduced into Japan in the 6th century AD. Shintō has no founder, no official sacred scriptures in the strict sense, and no fixed dogmas, but it has preserved its guiding beliefs throughout the ages.

This article is divided into the following sections:

Nature and varieties 278
History to 1900 278
- Early clan religion and ceremonies
- Early Chinese influences on Shintō
- The encounter with Buddhism
- Shintō reaction against Buddhism
- Neo-Confucian Shintō
- Fukko Shintō
- Formation of Sect Shintō

Shintō literature and mythology 280
Doctrines 280
- Concept of the sacred
- Precepts of truthfulness and purification
- Nature of man and other beliefs

Ritual practices and institutions 281
- Rites of passage
- Varieties of festival, worship, and prayer
- Types of shrines
- Other practices and institutions

Shintō religious arts 282
Political and social roles 282
Place of Shintō in Japanese and world religion 283
Bibliography 283

NATURE AND VARIETIES

Shintō consists of the traditional Japanese religious practices as well as the beliefs and life attitudes that are in accord with these practices. Shintō is more readily observed in the social life of the Japanese people and in their personal motivations than in a pattern of formal belief or philosophy. It remains closely connected with the Japanese value system and the Japanese people's ways of thinking and acting.

By courtesy of the Japan National Tourist Organization

Procession at Tōshō-gū (Tōshō Shrine) Festival at Nikkō, Japan, May 17–18.

Shintō can be roughly classified into the following three major types: Shrine Shintō, Sect Shintō, and Folk Shintō. Shrine Shintō (Jinja Shintō), which has been in existence from the beginning of Japanese history to the present day, constitutes a main current of Shintō tradition. Shrine Shintō includes within its structure the now defunct State Shintō (Kokka Shintō)—based on the total identity of religion and state—and has close relations with the Japanese Imperial family. Sect Shintō (Kyōha Shintō) is a relatively new movement consisting of 13 major sects that originated in Japan around the 19th century and of several others that emerged after World War II. Each sect was organized into a religious body by either a founder or a systematizer. Folk Shintō (Minzoku Shintō) is an aspect of Japanese folk belief that is closely connected with the other types of Shintō. It has no formal organizational structure nor doctrinal formulation but is centred in the veneration of small roadside images and in the agricultural rites of rural families. These three types of Shintō are interrelated: Folk Shintō exists as the substructure of Shintō faith, and a Sect Shintō follower is usually also a parishioner (*ujiko*) of a particular Shintō shrine.

Three major types of Shintō

HISTORY TO 1900

Much remains unknown about religion in Japan during the Paleolithic and Neolithic ages. It is unlikely, however, that the religion of these ages has any direct connection with Shintō. Yayoi culture, which originated in the northern area of the island of Kyushu in about the 3rd or 2nd century BC, is directly related to later Japanese culture and hence to Shintō. Among the primary Yayoi religious phenomena were agricultural rites and shamanism.

Early clan religion and ceremonies. In ancient times small states were gradually formed at various places. By the middle of the 4th century AD, a nation with an ancestor of the present Imperial Household as its head had probably been established. The constituent unit of society at that time was the *uji* (clan or family), and the head of each *uji* was in charge of worshiping the clan's *ujigami*—its particular tutelary or guardian deity. The prayer for good harvest in spring and the harvest ceremony in autumn were two major festivals honouring the *ujigami*. Divination, water purification, and lustration (ceremonial purification), which are all mentioned in the Japanese classics, became popular, and people started to build shrines for their *kami*.

Ancient Shintō was polytheistic. People found *kami* in nature, which ruled seas or mountains, as well as in outstanding men. They also believed in *kami* of ideas such as growth, creation, and judgment. Though each clan made the tutelary *kami* the core of its unity, such *kami* were not necessarily the ancestral deities of the clan. Sometimes *kami* of nature and *kami* of ideas were regarded as their tutelary *kami*.

Kami

Two different views of the world were present in ancient Shintō. One was the three-dimensional view in which the Plain of High Heaven (Takama no Hara, the *kami*'s world), Middle Land (Nakatsukuni, the present world), and the Hades (Yomi no Kuni, the world after death) were arranged in vertical order. The other view was a two-dimensional one in which this world and the Perpetual Country (Tokoyo, a utopian place far beyond the sea) existed in horizontal order. Though the three-dimensional view of the world (which is also characteristic of North Siberian and Mongolian shamanistic culture) became the representative view observed in Japanese myths, the two-dimensional view of the world (which is also present in Southeast Asian culture) was dominant among the populace.

Early Chinese influences on Shintō. Confucianism is believed to have reached Japan in the 5th century AD,

and by the 7th century it had spread among the people, together with Chinese Taoism and yin-yang (harmony of two basic forces of nature) philosophy. All of these stimulated the development of Shintō ethical teachings. With the gradual centralization of political power, Shintō began to develop as a national cult as well. Myths of various clans were combined and reorganized into a pan-Japanese mythology with the Imperial Household as its centre. The *kami* of the Imperial Household and the tutelary *kami* of powerful clans became the *kami* of the whole nation and people, and offerings were made by the state every year. Such practices were systematized supposedly around the start of the Taika-era reforms in 645. By the beginning of the 10th century, about 3,000 shrines throughout Japan were receiving state offerings. As the power of the central government declined, however, the system ceased to be effective, and after the 13th century only a limited number of important shrines continued to receive the Imperial offerings. Later, after the Meiji Restoration in 1868, the old system was revived.

Shintō–Buddhist syncretism

The encounter with Buddhism. Buddhism was officially introduced into Japan in AD 552 and developed gradually. In the 8th century there emerged tendencies to interpret Shintō from a Buddhist viewpoint. Shintō *kami* were viewed as protectors of Buddhism; hence shrines for tutelary *kami* were built within the precincts of Buddhist temples. *Kami* were made equivalent to *deva* (the Buddhist Sanskrit term for "gods") who rank highest in the Realm of Ignorance, according to Buddhist notions. Thus *kami*, like other creatures, were said to be suffering because they were unable to escape the endless cycle of transmigration; help was therefore offered to *kami* in the form of Buddhist discipline. Buddhist temples were even built within Shintō shrine precincts, and Buddhist sutras (scriptures) were read in front of *kami*. By the late 8th century *kami* were thought to be avatars, or incarnations, of buddhas and bodhisattvas. Bodhisattva names were given to *kami*, and Buddhist statues were placed even in the inner sanctuaries of Shintō shrines. In some cases, Buddhist priests were in charge of the management of Shintō shrines.

From the beginning of the Kamakura period (1192–1333), theories of Shintō-Buddhist amalgamation were formulated. The most important of the syncretic schools to emerge were Ryōbu (Dual Aspect) Shintō and Sannō ("King of the Mountain," a common name of the guardian deity of Tendai Buddhism) Shintō. According to Ryōbu Shintō—also called Shingon Shintō—the two realms of the universe in Shingon Buddhist teachings corresponded to the *kami* Amaterasu Ōmikami and Toyuke (Toyouke) Ōkami enshrined at the Ise-daijingū (Grand Shrine of Ise, commonly called Ise-jingū, or Ise Shrine) in Mie prefecture. The theorists of Sannō Shintō—also called Tendai Shintō—interpreted the Tendai belief in the central, or absolute, truth of the universe (*i.e.*, the fundamental buddha nature) as being equivalent to the Shintō concept that the sun goddess Amaterasu was the source of the universe. These two sects brought certain esoteric Buddhist rituals into Shintō. Buddhistic Shintō was popular for several centuries and was influential until its extinction at the Meiji Restoration.

Shintō reaction against Buddhism. Ise, or Watarai, Shintō was the first theoretical school of anti-Buddhistic Shintō in that it attempted to exclude Buddhist accretions and also tried to formulate a pure Japanese version. Watarai Shintō appeared in Ise during the 13th century as a reaction against the Shintō-Buddhist amalgamation. Konton (chaos), or Kizen (non-being), was the basic *kami* of the universe for Watarai Shintō and was regarded as the basis of all beings, including the buddhas and bodhisattvas. Purification, which had been practiced since the time of ancient Shintō, was given much deeper spiritual meanings. *Shōjiki* (defined as uprightness or righteousness) and prayers were emphasized as the means by which to be united with *kami*.

Yoshida Shintō, a school in Kyōto that emerged during the 15th century, inherited various aspects handed down from Watarai Shintō and also showed some Taoist influence. The school's doctrines were largely the work of Yoshida Kanetomo (1435–1511). Its fundamental *kami* (the source of all things and beings in the universe) was Taigen Sonjin (the Great Exalted One). According to its teaching, if one is truly purified, his heart can be the *kami*'s abode. The ideal of inner purification was a mysterious state of mind in which one worshiped the *kami* that lived in one's own heart. Although the Watarai and Yoshida schools were thus free of Buddhistic theories, the influence of Chinese thought was still present.

Neo-Confucian Shintō. In 1603 the Tokugawa shogunate was founded in Edo (Tokyo), and contact between Shintō and Confucianism was resumed. Scholars tried to interpret Shintō from the standpoint of Neo-Confucianism, emphasizing the unity of Shintō and Confucian teachings. Schools emerged based on the teachings of the Chinese philosophers Chu Hsi and Wang Yang-ming, and Neo-Confucianism became an official subject of study for warriors. Yoshikawa Koretaru (1616–94) and Yamazaki Ansai (1619–82) were two representative scholars of Confucian Shintō. They added Neo-Confucian interpretations to the traditional theories handed down from Watarai Shintō, and each established a new school. The T'ai Chi (Supreme Ultimate) concept of Neo-Confucianism was regarded as identical with the first *kami* of the *Nihon shoki*, or *Nihon-gi* ("Chronicles of Japan"). One of the characteristics of Yoshikawa's theories was his emphasis on political philosophy. Imperial virtues (wisdom, benevolence, and courage), symbolized by the Sanshu no Shinki (Three Sacred Treasures), and national ethics, such as loyalty and filial piety, constituted the way to rule the state. Yamazaki Ansai further developed this tendency and advocated both mystic pietism and ardent emperor worship.

Yoshikawa Koretaru and Yamazaki Ansai

Fukko Shintō. Fukko (Restoration, or Revival) Shintō is one of the Kokugaku (National Learning) movements that started toward the end of the 17th century. Advocates of this school maintained that the norms of Shintō should not be sought in Buddhist or Confucian interpretations but in the beliefs and life-attitudes of their ancestors as clarified by philological study of the Japanese classics. Motoori Norinaga (1730–1801) represented this school. His emphasis was on the belief in *musubi* (the mystical power of becoming or of creation), which had been popular in ancient Shintō, and on a this-worldly view of life, which anticipated the eternal progress of the world in ever-changing mutations. These beliefs, together with the inculcation of respect for the Imperial line and the teaching of absolute faith—according to which all problems beyond human capability were turned over to *kami*—exercised great influence on modern Shintō doctrines.

The most important successor of Motoori in the field of Shintō was Hirata Atsutane (1776–1843), who showed the influence of Roman Catholic teachings in some respects—derived from the writings of Jesuits in China—by advancing the idea of a creator god and retribution for ethical and religious failings in another world. These doctrines, however, were not accepted into the main current of Shintō. Hirata developed the philological studies started by Motoori and trained many capable disciples. He also wrote prayers, worked out formulas for family cults of tutelary *kami* and ancestors, and promoted Shintō practices. His spirituality, reverence for the emperor, and desire to restore the spirit of ancient Shintō enlisted many supporters and served as one of the factors in bringing about the Meiji Restoration in 1868.

Formation of Sect Shintō. During the latter part of the 19th century, new religious movements emerged out of the social confusion and unrest of the people. What these new movements taught differed widely: some were based on mountain-worship groups, which were half Buddhist and half Shintō; some placed emphasis on purification and ascetic practices; and some combined Confucian and Shintō teachings. New religious movements—such as Kurozumi-kyō (in this sense *kyō* means "religion," or "religious body"), founded by Kurozumi Munetada (1780–1850); Konkō-kyō (Konkō is the religious name of the founder of this group and means, literally, "golden light") by Kawate Bunjirō (1814–83); and Tenri-kyō (*tenri* means "divine reason or wisdom") by Nakayama Miki (1798–1887)—were based mostly on individual religious experiences and aimed at healing diseases or spiritual salvation.

These sectarian Shintō groups, numbering 13 during the Meiji period (1868–1912), were stimulated and influenced by Restoration Shintō. They can be classified as follows:

1. Revival Shintō sects: Izumo-ōyashiro-kyō (or Taisha-kyō), Shintō-taikyō, Shinri-kyō
2. Confucian sects: Shintō Shūsei-ha, Shintō Taisei-kyō
3. Purification sects: Shinshū-kyō, Misogi-kyō
4. Mountain worship sects: Jikkō-kyō, Fusō-kyō, On take-kyō (or Mitake-kyō)
5. "Faith-healing" sects: Kurozumi-kyō, Konkō-kyō, Tenri-kyō

SHINTŌ LITERATURE AND MYTHOLOGY

Kojiki and *Nihon-gi* (*Nihon shoki*)

Broadly speaking, Shintō has no founder. When the Japanese people and Japanese culture became aware of themselves, Shintō was already there. Nor has it any official scripture that can be compared to the Bible in Christianity or to the Qur'ān in Islām. The *Kojiki* ("Records of Ancient Matters") and the *Nihon-gi,* or *Nihon shoki* ("Chronicles of Japan"), are regarded in a sense as sacred books of Shintō. They were written in AD 712 and 720, respectively, and are compilations of the oral traditions of ancient Shintō. But they are also books about the history, topography, and literature of ancient Japan. It is possible to construct Shintō doctrines from them by interpreting the myths and religious practices they describe.

Stories partially similar to those found in Japanese mythology can be found in the myths of Southeast Asia; and in the style of description in Japanese myths some Chinese influence is detectable. The core of the mythology, however, consists of tales about the sun goddess Amaterasu Ōmikami, the ancestress of the Imperial Household, and tales of how her direct descendants unified the Japanese people under their authority. In the beginning, according to Japanese mythology, a certain number of *kami* simply emerged, and a pair of *kami,* Izanagi and Izanami, gave birth to the Japanese islands, as well as to the *kami* who became ancestors of the various clans. Amaterasu, the ruler of Takama no Hara; the moon god Tsukiyomi no Mikoto; and Susanoo (Susanowo) no Mikoto, the ruler of the nether regions, were the most important among them. A descendant of Amaterasu, Jimmu, is said to have become the first emperor of Japan. Japanese mythology says that the Three Sacred Treasures (the mirror, the sword, and the jewels), which are still the most revered symbols of the Imperial Household, were first given by Amaterasu to her grandson. The Inner Shrine (Naikū) of the Isejingū is dedicated to this ancestral goddess and is the most venerated shrine in Shintō.

The Japanese classics also contain myths and legends concerning the so-called 800 myriads of *kami* (*yao-yorozu no kami;* literally, *yao* equals 800 and *yorozu* 10,000). Some of them are the tutelary deities of clans and later became the tutelary *kami* of their respective local communities. Many others, however, are not enshrined in sanctuaries and have no direct connections with the actual Shintō faith.

DOCTRINES

Kami and *makoto*

Concept of the sacred. At the core of Shintō are beliefs in the mysterious creating and harmonizing power (*musubi*) of *kami* and in the truthful way or will (*makoto*) of *kami.* The nature of *kami* cannot be fully explained in words, because *kami* transcends the cognitive faculty of man. Devoted followers, however, are able to understand *kami* through faith and usually recognize various *kami* in polytheistic form.

Parishioners of a shrine believe in their tutelary *kami* as the source of human life and existence. Each *kami* has a divine personality and responds to truthful prayers. The *kami* also reveals *makoto* to people and guides them to live in accordance with it. In traditional Japanese thought, truth manifests itself in empirical existence and undergoes transformation in infinite varieties in time and space. *Makoto* is not an abstract ideology. It can be recognized every moment in every individual thing in the encounter between man and *kami.*

In Shintō all the deities are said to cooperate with one another, and life lived in accordance with a *kami*'s will is believed to produce a mystical power that gains the protection, cooperation, and approval of all the particular *kami.*

Precepts of truthfulness and purification. As the basic attitude toward life, Shintō emphasizes *makoto no kokoro* ("heart of truth"), or *magokoro* ("true heart"), which is usually translated as "sincerity, pure heart, uprightness." This attitude follows from the revelation of the truthfulness of *kami* in man. It is, generally, the sincere attitude of a person in doing his best in the work he has chosen or in his relationship with others, and the ultimate source of such a life-attitude lies in man's awareness of the divine.

Although Shintō ethics do not ignore individual moral virtues such as loyalty, filial piety, love, faithfulness, and so forth, it is generally considered more important to seek *magokoro,* which constitutes the dynamic life-attitude that brings forth these virtues. In ancient scriptures *magokoro* was interpreted as "bright and pure mind" or "bright, pure, upright, and sincere mind." Purification, both physical and spiritual, is stressed even in contemporary Shintō to produce such a state of mind. The achievement of this state of mind is necessary in order to make communion between *kami* and man possible and to enable individuals to accept the blessings of *kami.*

Nature of man and other beliefs. In Shintō it is commonly said that "man is *kami*'s child." First, this means that a person was given his life by *kami* and that his nature is therefore sacred. Second, it means that daily life is made possible by *kami,* and, accordingly, the personality and life of people are worthy of respect. An individual must revere the basic human rights of everyone (regardless of race, nationality, and other distinctions) as well as his own. The concept of original sin is not found in Shintō. On the contrary, man is considered to have a primarily divine nature. In actuality, however, this sacred nature is seldom revealed in man. Purification is considered symbolically to remove the dust and impurities that cover one's inner mind.

Role of the individual in the history of his family

Shintō is described as a religion of *tsunagari* ("continuity or communion"). The Japanese, while recognizing each man as an individual personality, do not take him as a solitary being separated from others. On the contrary, he is regarded as the bearer of a long, continuous history that comes down from his ancestors and continues in his descendants. He is also considered as a responsible constituent of various social groups.

Motoori Norinaga stated that the human world keeps growing and developing while continuously changing. Similarly, Japanese mythology speaks of an eternity of history in the divine edict of Amaterasu. In its view of history, Shintō adheres to the cyclical approach, according to which there is a constant recurrence of historical patterns. Shintō does not have the concept of the "last day": there is no end of the world or of history. One of the divine edicts of Amaterasu says:

> This Reed-plain-1,500-autumns-fair-rice-ear Land is the region which my descendants shall be lords of. Do thou, my August Grandchild, proceed thither and govern it. Go! and may prosperity attend thy dynasty, and may it, like Heaven and Earth, endure forever.

Modern Shintōists interpret this edict as revealing the eternal development of history as well as the eternity of the dynasty. From the viewpoint of finite individuals, Shintōists also stress *naka-ima* ("middle present"), which repeatedly appears in the Imperial edicts of the 8th century. According to this point of view, the present moment is the very centre in the middle of all conceivable times. In order to participate directly in the eternal development of the world, it is required of Shintōists to live fully each moment of life, making it as worthy as possible.

Historically, the *ujigami* of each local community played an important role in combining and harmonizing different elements and powers. The Imperial system, which has been supported by the Shintō political philosophy, is an example of unity and harmony assuming the highest cultural and social position in the nation. After the Meiji Restoration (1868), Shintō was used as a means of spiritually unifying the people during repeated wars. Since the end of World War II, the age-old desire for peace has been reemphasized. *The General Principles of Shintō*

Life proclaimed by the Association of Shintō Shrines in 1956 has the following article: "In accordance with the Emperor's will, let us be harmonious and peaceful, and pray for the nation's development as well as the world's co-prosperity."

RITUAL PRACTICES AND INSTITUTIONS

Shintō does not have a weekly religious service. People visit shrines at their convenience. Some may go to the shrines on the 1st and 15th of each month and on the occasions of rites or festivals (*matsuri*), which take place several times a year. Devotees, however, may pay respect to the shrine every morning.

Matsuri

Rites of passage. Various Shintō rites of passage are observed in Japan. The first visit of a newborn baby to the tutelary *kami*, which occurs 30 to 100 days after birth, is to initiate the baby as a new adherent. The Shichi-go-san (Seven-Five-Three) festival on November 15 is the occasion for boys of five years and girls of three and seven years of age to visit the shrine to give thanks for *kami*'s protection and to pray for their healthy growth. January 15 is Adults' Day. Youth in the village used to join the local young men's association on this day. At present it is the commemoration day for those Japanese who have attained their 20th year. The Japanese usually have their wedding ceremonies in Shintō style and pronounce their wedding vows to *kami*. Shintō funeral ceremonies, however, are not popular. The majority of the Japanese are Buddhist and Shintōist at the same time and have their funerals in Buddhist style. A traditional Japanese house has two family altars: one, Shintō, for their tutelary *kami* and the goddess Amaterasu Ōmikami, and another, Buddhist, for the family ancestors. Pure Shintō families, however, will have all ceremonies and services in Shintō style. There are other Shintō *matsuri* concerning occupations or daily life, such as a ceremony of purifying a building site or for setting up the framework for a new building, a firing or purifying ceremony for the boilers in a new factory, a completion ceremony for a construction works, or a launching ceremony for a new ship.

Varieties of festival, worship, and prayer. Each Shintō shrine has several major festivals each year, including the Spring Festival (Haru Matsuri, or Toshigoi-no-Matsuri; Prayer for Good Harvest Festival), Autumn Festival (Aki Matsuri, or Niiname-sai; Harvest Festival), an Annual Festival (Rei-sai), and the Divine Procession (Shinkō-sai). The Divine Procession usually takes place on the day of the Annual Festival, and miniature shrines (*mikoshi*) carried on the shoulders are transported through the parish. The order of rituals at a grand festival is usually as follows:

1. Purification rites (*harae*)—commonly held at a corner of the shrine precincts before participants come into the shrine but sometimes held within the shrine before beginning a ceremony.
2. Adoration—the chief priest and all the congregation bow to the altar.
3. Opening of the door of the inner sanctuary (by the chief priest).
4. Presentation of food offerings—rice, sake wine, rice cakes, fish, seaweed, vegetables, salt, water, etc., are offered but animal meat is not, because of the taboo on shedding blood in the sacred area. In the past cooked food was usually offered to *kami*, but nowadays uncooked food is more often used. In accordance with this change, the idea of entertaining *kami* changed to that of thanksgiving.
5. Prayer—the chief priest recites prayers (*norito*) modeled on ancient Shintō prayers. These prayers were compiled in the early 10th century and were based on the old belief that spoken words had spiritual potency.
6. Sacred music and dance.
7. General offering—participants in the festival make symbolic offerings using little branches of the evergreen sacred tree to which strips of white paper are tied.
8. Taking offerings away.
9. Shutting the door of the inner sanctuary.
10. Final adoration.
11. Feast (*naorai*).

Ritual at festivals

Inner sanctuary (centre background) of the Grand Shrine of Izumo (Izumo-taisha), dedicated to Ōkuninushi no Kami.
By courtesy of the Izumo-taisha

In the olden days *naorai*, a symbolic action in which participants held communion with *kami* by having the same food offered to the deity, came in the middle of the festival ceremony. The custom is still observed sometimes at the Imperial Household and at some old shrines, but it is more common to have communion with *kami* by drinking the offered sake after the festival. Since World War II it has become popular to have a brief sermon or speech before the feast.

Most Shintō festivals are observed generally in accordance with the above-mentioned order. On such occasions as the Annual Festival, various special rites may be held—for example, special water purification (*misogi*) and confinement in shrines for devotional purposes (*o-komori*), the procession of a sacred palanquin (*o-miyuki*) or of boats (*funa matsuri*), a ceremonial feast (*tōya matsuri*), sumo wrestling, horseback riding (*kurabe-uma*), archery (*matoi*), a lion dance (*shishi mai*), and a rice-planting festival (*o-taue matsuri*).

Types of shrines. A simple torii (gateway) stands at the entrance of the shrine precincts (see Tōshō-gū illustration above). After proceeding on the main approach, a visitor will come to an ablution basin where the hands are washed and the mouth is rinsed. Usually he will make a small offering at the oratory (*haiden*) and pray. Sometimes a visitor may ask the priest to conduct rites of passage or to offer special prayers. The most important shrine building is the main, or inner, sanctuary (*honden*), in which a sacred symbol called *shintai* ("*kami* body") or *mitama-shiro* ("divine spirit's symbol") is enshrined. The usual symbol is a mirror, but sometimes it is a wooden image, a sword, or some other object. In any case, it is carefully wrapped and placed in a container. It is forbidden to see it: only the chief priest is allowed to enter inside the inner sanctuary.

Shintai and *mitama-shiro*

In the beginning Shintō had no shrine buildings. At each festival people placed a tree symbol at a sacred site, or they built a temporary shrine to invite *kami*. Later they began to construct permanent shrines where *kami* were said to stay permanently. The *honden* of the Inner Shrine at Ise and of Izumo-taisha (Grand Shrine of Izumo, in Shimane prefecture) illustrate two representative archetypes of shrine construction. The style of the former probably developed from that of a storehouse for crops, especially for rice, and the style of the latter from ancient house construction. In the course of time, variations of shrine architecture were adopted and additional buildings were attached in front of the *honden*. The *honden* and *haiden* are in many cases connected by a hall of offering (*heiden*) where prayers are usually recited. Large shrines also have a hall for liturgical dancing (*kaguraden*).

Other practices and institutions. *Ujigami* belief is the most popular form of Shintō in Japan. Originally referring to the *kami* of an ancient clan, after the 13th century *ujigami* was used in the sense of the tutelary *kami* of a

local community, and all the members in the community were that *kami*'s adherents (*ujiko*). Even today a *ujiko* group consists of the majority of the residents in a given community. A Shintōist, however, can believe at the same time in shrines other than his own local shrine. It was only after World War II that some large shrines also started to organize believers' groups (*sūkeisha*). The Believers' Association of the Meiji Shrine, for instance, has about 240,000 members living in and around Tokyo.

Kokugakuin University in Tokyo and Kōgakkan University at Ise are the primary training centres for Shintō priests. Though any Shintōists who go through certain training processes may be a priest (or a priestess), many priests are, in fact, from the families of hereditary Shintō priests.

SHINTŌ RELIGIOUS ARTS

The Japanese from ancient times have valued emotional and aesthetic intuitions in expressing and appreciating their religious experiences. They found symbols of *kami* in natural beauty and the forces of nature, and they developed explicitly religious poetry, architecture, and visual arts. Shrine precincts are covered with green trees and are places of a serene and solemn atmosphere, which is effective in calming worshipers' minds. In the larger shrines, surrounded by expansive woods with mountains as their background, a harmony of nature and architecture may be achieved. Ise-jingū and Izumo-taisha still retain the ancient architectural styles. After the 9th century an intricate form of shrine construction was developed, adopting both Buddhist and Chinese architectural styles and techniques. The curving roof style is one example. Unpainted timbers are most frequently used, but, wherever Buddhistic Shintō was popular, Chinese vermilion-lacquered shrines were also built.

Torii: the sacred gates

A torii always stands in front of a shrine. Various kinds of torii can be seen in Japan, but their function is always the same: to divide the sacred precincts from the secular area. A pair of sacred stone animals called *komainu* ("Korean dogs") or *karajishi* ("Chinese lions") are placed in front of a shrine. Originally they served to protect the sacred buildings from evil and defilements. After the 9th century they were used for ornamental purposes on ceremonial occasions at the Imperial Court and later came to be used at various shrines generally. Some of the stone lanterns (*ishidōrō*) used at the shrines are works of art. The dedicator's name and the year are inscribed on the lanterns to inform viewers of the long tradition of faith and to urge them to maintain it.

Compared with Buddhist statuary, visual representations of *kami* are not outstanding either in their quality or quantity. Images of *kami* were, in fact, not used in ancient Shintō until after the introduction of Buddhism into Japan. These are placed in the innermost part of the *honden* and are not the objects of direct worship by the people. *Kami* icons are not worshiped at shrines.

The history of the shrine, its construction arrangements, and ritual processions are recorded in picture scrolls (*emakimono*), and at the older shrines there are many votive pictures (*ema*)—small wooden picture plaques—that have been dedicated over the years by worshipers. Other articles, such as specimens of calligraphy, sculpture, swords, and arms, dedicated by the Imperial families, nobles, or feudal lords, are also kept at shrines. Several hundred such items and shrine constructions have been designated by the Japanese government as national treasures and important cultural properties.

Religious music and dance

The traditional religious music and dance of shrines were performed for the purpose of entertaining and appeasing *kami*, rather than to praise them. *Gagaku* (literally, "elegant music") involves both vocal and instrumental music, specifically for wind, percussion, and stringed instruments. *Gagaku* with dance is called *bugaku*. *Gagaku* was patronized by the Imperial Household as court music and was much appreciated by the upper classes from the 9th to the 11th century. Later some of the more solemn and graceful pieces were used as ritualistic music by shrines and temples. Today *gagaku* is widely performed at larger shrines. The authentic tradition of *gagaku* has been transmitted by the Bureau of Music (Gagaku-ryō, now called Gakubu) of the Imperial Household (established in 701).

Apart from *gagaku* there are also *kagura* (a form of indigenous religious music and dance based on blessing and purification), *ta-asobi* (a New Year's dance-pantomime of the cycle of rice cultivation), and *shishi mai*, which developed originally from magico-religious dances and are now danced for purification and as prayers. *Matsuri-bayashi* is a gay, lively music with flutes and drums to accompany divine processions. Some organizations of both Shrine and Sect Shintō have recently begun to compose solemn religious songs to praise *kami*, making use of Western musical forms.

POLITICAL AND SOCIAL ROLES

Until the end of World War II, Shintō was closely related to the state. Offerings to *kami* were made every year by the government and the Imperial Household, and prayers were offered for the safety of the state and people. The *matsuri-goto* (the affairs of worship) offered by the emperor from olden days included not only ceremonies for *kami* but also for ordinary matters of state. "Shintō ceremonies and political affairs are one and the same" was the motto of officials. Administrators were required to have a religious conscience and develop political activities with *magokoro*.

This tradition was maintained as an undercurrent throughout Japanese history. Villagers prayed to the tutelary *kami* of the community for their peace and welfare and promoted unity among themselves with village festivals. After the Meiji Restoration, the government treated Shintō like a state religion and revived the system of national shrines, which dated from the 9th century or earlier. In order to propagate Revival Shintō as the foundation of the national structure, they initiated the "great promulgation movement" (1869–84) in which the emperor was respected like *kami*. Although the Japanese constitution enacted in 1889 guaranteed freedom of faith under certain conditions, priority was, in fact, given to Shintō. In elementary schools Shintō was taught to children, and most of the national holidays were related to Shintō festivals. Shintō of this nature was called State Shintō and came under the control of the Bureau of Shrines in the Ministry of Home Affairs.

Relation of State and Sect Shintō

State Shintō was regarded as a state cult and a national ethic and not as "a religion." The free interpretation of its teachings by individual Shintō priests was discouraged. Priests of the national shrines were prohibited from preaching and presiding over Shintō funerals. By 1945 there were 218 national and approximately 110,000 local shrines. The number of Sect Shintō groups was limited to 13 after the organization of Tenri-kyō. Legally these 13 sects were treated as general religious bodies, similar to Buddhism and Christianity, and came under the supervision of the Ministry of Education.

After the end of World War II, the Supreme Commander for the Allied Powers ordered the Japanese government to disestablish State Shintō. All government financial support from public funds and all official affiliation with Shintō and Shintō shrines were also discontinued. State rites performed by the emperor were henceforth to be regarded as the religious practices of the Imperial family. These rulings were carried into the new Japanese constitution that was enacted in 1947. Presently, Shrine Shintō is faced with two serious problems. The first is determining how the traditional unifying function of Shintō can be promoted in local communities or in the nation without interfering with freedom of faith. The second is the necessity of harmonizing Shintō with rapid modernization, especially in organizing believers and dealing with human problems or the meaning of life.

The number of Shintō shrines has been decreasing since the beginning of the Meiji era, in part because a municipal unification plan in 1889 called for the shrines of tutelary *kami* to be combined with the municipality. At present, about 99 percent of the shrines belong to the Association of Shintō Shrines, established in 1946, and most of the others are independent or belong to small groups.

About 15 percent of 16,251 Sect Shintō churches were

damaged during World War II. Although they were not affected by the occupation policies after the war, many sects, in fact, went through difficult years because of unrest among the people and disunion within their own organizations. In 1966 Tenri-kyō proclaimed that their belief was not Shintō, and in 1973 they withdrew from the federation of Sect Shintō groups. On the other hand, numerous new religious bodies, including Shintōist groups, have emerged since 1945. How to adequately reclassify Sect Shintō, when combined with these new bodies, is a major concern of specialists on the subject.

PLACE OF SHINTŌ IN JAPANESE AND WORLD RELIGION

Shintō together with Buddhism is closely related both culturally and socially to the life of the Japanese people. Its relationships to other religions in Japan are generally cooperative and harmonious. Most Shintōists believe that cooperation between different religions could contribute to world peace, but this is not to imply a facile religious syncretism. Shintōists insist on maintaining their own characteristics and inner depth while working toward the peaceful coexistence of human beings.

BIBLIOGRAPHY. H. BYRON EARHART, *Japanese Religion: Unity and Diversity*, 3rd ed. (1982), examines the formation, development, and interaction of religions. JOSEPH M. KITAGAWA, *Religion in Japanese History* (1966, reissued 1990), is a widely used survey textbook on Japanese religious background. Studies specifically about Shintō include NAOFUSA HIRAI (HIRAI NAOFUSA), *Japanese Shinto* (1966), a brief general sketch; STUART D.B. PICKEN, *Shinto: Japan's Spiritual Roots* (1980), a short introduction to the origins and modern forms of Shintō; D.C. HOLTOM, *The National Faith of Japan* (1938, reprinted 1965), strong in history and political philosophy; TSUNETSUGU MURAOKA, *Studies in Shinto Thought* (1964, reprinted 1988), a dependable description of Shintō thought by an eminent philologist; and ICHIRŌ HORI, *Folk Religion in Japan: Continuity and Change* (1968, reprinted 1983), a good study on the religious and social background of folk Shintō. ROBERT S. ELLWOOD, *The Feast of Kingship* (1973), describes the ancient enthronement ceremonies of Japanese emperors. Editions of the sacred books include W.G. ASTON (trans.), *Nihongi: Chronicles of Japan from the Earliest Times to A.D. 697*, 2 vol. (1896, reissued 2 vol. in 1, 1972), a standard translation into English; and DONALD L. PHILIPPI (trans.), *Kojiki* (1968, reissued 1992), a translation with introduction using contemporary Japanese philological studies.

(N.H.)

Sikhism

Sikhism is the religion of an Indian group founded in the Punjab (or Pañjāb) in the late 15th century AD by Gurū Nānak. Its members are known as Sikhs.

In the early 21st century there were more than 24 million Sikhs in the world. The great majority of Sikhs live in the state of Punjab. Most of the remainder are in Haryāna state and Delhi or are scattered in other parts of India. Some Sikhs have also settled in Malaysia, Singapore, East Africa, the United Kingdom, the United States, and Canada. The word Sikh is derived from the Pāli *sikkha* or Sanskrit *śiṣya,* meaning "disciple." Sikhs are disciples of their Ten Gurūs (religious teachers), beginning with Nānak (1469–1539) and ending with Gobind Singh (1666–1708).

This article is divided into the following sections:

History and background 284
Religious and cultural origins
The Ten Gurūs: Nānak and his tradition
The founding of the Khālsā
The Sikh empire of Ranjit Singh
Relations between the Sikhs and the British
Sikhism since 1947
Sikh literature, myth, and lore 286
Canonical and noncanonical literature
Myths and lore
Sikh doctrines, practices, and institutions 286
Doctrines
Practices and institutions
Bibliography 287

HISTORY AND BACKGROUND

Religious and cultural origins. Sikhism was a historical development of the Hindu Vaiṣṇava Bhakti movement—a devotional movement among followers of the god Vishnu—that began in Tamil country and was introduced to the north by Rāmānuja (traditionally, 1017–1137). In the 14th and 15th centuries, and after prolonged confrontation with Islām, the movement spread across the Indo-Gangetic Plain. The Bhaktas (devotees) maintained that God, though known by many names and beyond comprehension, is the one and the only reality; that all else is illusion (*māyā*); and that the best way to approach God is through repetition of his name (Sanskrit *nāma*), singing hymns of praise (Punjabi *kīrtan*), and meditation under the guidance of a Gurū. Traditional Hindu religion and society were hierarchically structured; the Bhakti movement opposed the Brahmin hegemony over religious ritual and the caste system.

Hindu Vaiṣṇava and Islāmic Ṣūfī origins

Kabīr (1440–1518), a mystic poet and religious synthesist, was the link between Hindu Bhakti and Islāmic Ṣūfism, which had gained a large following among Indian Muslims. Ṣūfīs (mystics) also believed in singing hymns and in meditation under the guidance of a leader. They welcomed non-Muslims in their hospices. Sikhism drew inspiration from both Bhaktas and Ṣūfīs.

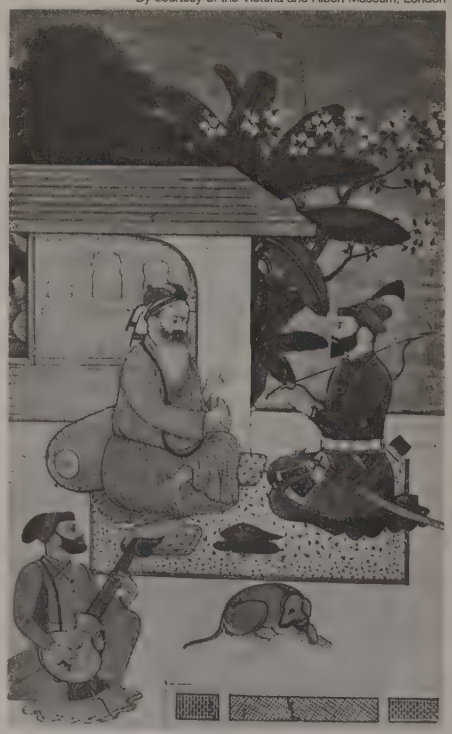
The Ten Gurūs: Nānak and his tradition. Nānak was born in 1469 in the village of Rāi Bhoi dī Talvaṇḍī, 40 miles (65 kilometres) from Lahore (in present-day Pakistan). His father was a revenue collector belonging to the Bedī (conversant with the Vedas—the revealed scriptures of Hinduism) subcaste of Kṣatriyas (Warriors). Nānak received an education in traditional Hindu lore and in the rudiments of Islām. Early in life he began associating with holy men. For a time he worked as the accountant of the Afghān chieftain at Sultānpur. There a Muslim family servant, Mardānā, who was also a rebec player, joined him. Nānak began to compose hymns. Mardānā put them to music and the two organized community hymn singing. They organized a *laṅgar* ("free community kitchen") where Muslims, as well as Hindus of different castes, could eat together. At Sultānpur, Nānak had his first vision of God, in which he was ordered to preach to mankind. He disappeared while bathing in a stream. When he reappeared on the third day, he proclaimed: "There is no Hindu, there is no Muslim."

Sikh tradition relates that Nānak also undertook four long voyages: east as far as Assam; south through the Tamil country to Sri Lanka; north to Ladakh and Tibet; and west as far as Mecca, Medina, and Baghdad. He spent the last years of his life in Kartārpur (in present-day Pakistan), where he raised the first Sikh temple. He nominated one of his disciples, Aṅgad, as his successor.

Aṅgad (Gurū 1539–52) was followed by another disciple, Amar Dās (Gurū 1552–74), who later nominated his son-in-law, Rām Dās Soḍhī (Gurū 1574–81), as his successor. Thereafter the office of Gurū remained in the Soḍhī family. Rām Dās was succeeded by his youngest son, Arjun Mal (Gurū 1581–1606), who, before his death by torture in Lahore on May 30, 1606, nominated his son Hargobind (Gurū 1606–44). The seventh Gurū, Har Rāī (Gurū 1644–61), was Hargobind's grandson, who, after his tenure, nominated his young son Hari Krishen (Gurū 1661–64), who died of smallpox at the age of eight. Tegh Bahādur (Gurū 1664–75), who succeeded him, was the son of the sixth Gurū, Hargobind. Before his execution in Delhi on November 11, 1675, Tegh Bahādur passed succession to his son, Gobind Rāī (Gurū 1675–1708).

The founding of the Khālsā. The execution of two Gurūs and persecution by the Mughals compelled the Sikhs to take to arms. This was given religious sanction when, on the Hindu New Year's Day (April 13, 1699), Gobind Rāī baptized five Sikhs into a new fraternity he called the Khālsā, meaning the "Pure" (from the Persian *khāleṣ,* also

By courtesy of the Victoria and Albert Museum, London



The first Sikh Gurū, Nānak, conversing with the 10th and last Gurū, Gobind Singh. The imaginary meeting is expressive of the religion's development from a pacifist to a militant brotherhood. Painting of the Guler school, *c.* 1820. In the collection of Mohan Singh, Punjab, India.

The surname Singh

meaning "pure"), and gave himself and them a common surname, Singh ("Lion"). Kaur ("Princess") is the corresponding name given to all Sikh women. As military leader of the Sikhs, Gobind Singh fought to protect the community against religious persecution. Facing opposition from the local Hindu chiefs, who were supported by the Mughal emperor Aurangzeb, Gobind Singh was forced to leave Anandpur in 1704. In the fighting that followed, two of his sons were killed and two more were captured and executed. On better terms with Aurangzeb's successor, Gobind Singh joined the new emperor on military campaigns but was assassinated at Nānded (now in Maharāshtra) on October 7, 1708. Before his death he declared that he was the last of the personal Gurūs and that the authority of Gurū would reside in the teachings of the *Ādi Granth,* the Sikhs' holy scriptures. The military leadership of the Sikhs devolved upon Bandā Singh Bahādur. For eight years Bandā defied the Mughals and devastated large tracts of east-central Punjab, until he was captured and, along with 700 of his followers, executed in Delhi in the summer of 1716.

For a few years the Khālsā disappeared into the hills. But when Mughal power was weakened by the incursion in 1738–39 of the Persian Nāder Shāh, they reemerged into the plains. They organized themselves under *misl*s (from Persian *mēsāl,* meaning both "example" and "equal") and began to extract protection money from towns and villages. The series of invasions between 1747 and 1769 that were led by Aḥmad Shāh Durrānī completely disrupted Mughal administration. In the battle of Panipat in 1761, the Afghāns destroyed rising Marāthā power in the north. The power vacuum thus created allowed the Sikhs to establish themselves as the rulers of the Punjab.

The Sikh empire of Ranjit Singh. In the years of turmoil between the Persian and Afghān invasions, Sikh *misl*s had operated in loosely defined areas. Two main divisions emerged: Cis-Sutlej, the area between the Sutlej and the Yamuna rivers, and Trans-Sutlej, between the Sutlej and the Indus rivers. In 1761 Sikhs wrested the capital city of Lahore from the Mughal governor.

Ranjit Singh

Ranjit Singh's (1780–1839) *misl,* the Śukerchakīās, was based at Gujrānwāla, north of Lahore. Ranjit took possession of the capital in 1799 and two years later had himself crowned maharaja of the Punjab. The English, who had advanced beyond Delhi, took the Cis-Sutlej states under their protection and compelled Ranjit Singh to accept the Sutlej River as the southeastern limit of his kingdom. There, after Ranjit Singh systematically brought the Trans-Sutlej region under his suzerainty, he took Multan in 1818 and Kashmir in 1819. In the following winter he extended his domain north and west beyond the Indus River into the land of the Pathans. In so doing, he sought to bring about the Khālsā Rāj, or kingdom of God on earth.

Ranjit Singh then began modernizing his army by employing European officers to train his troops. This army defeated the Pathans and Afghāns and extended Sikh power to the Khyber Pass. He also forbade capital punishment, offered large amounts to charity, and enhanced the beauty of the Golden Temple in Amritsar.

Relations between the Sikhs and the British. After taking the Cis-Sutlej states under their protection, the British began to make plans for extending their empire up to the Indus River. Even during Ranjit Singh's lifetime they had been interfering in the affairs of Afghanistan, and they had persuaded him to join in an Anglo-Sikh expedition to Kabul. After the death of Ranjit Singh, the Sikh kingdom disintegrated rapidly. Ranjit's eldest son and successor, Kharak Singh, was deposed by his own son, Naonihal Singh, and died of excessive use of opium. On the same day, Naonihal Singh was mortally injured when a gateway collapsed on his head. Kharak Singh's widow, Chand Kaur, occupied the throne for a few months until she was deposed and later murdered by Ranjit Singh's second son, Sher Singh. On September 15, 1843, Sher Singh, his son Pratap Singh, and Chief Minister Dhian Singh Dogra were murdered by Chand Kaur's kinsmen, who in their turn were slain by Dhian Singh's son, Hira Singh Dogra. Ranjit Singh's youngest son, Dalip Singh, was proclaimed maharaja with his mother, Jindan Kaur, as regent and Hira Singh Dogra as prime minister. Power passed, however, into the hands of the *pañcāyat* (elected council) of the Khālsā army, which compelled the Dogra to flee Lahore and then slew him in flight.

The British began to move their troops to the Sikh frontier and made preparations to cross the Sutlej. On December 11, 1845, the Khālsā army began crossing the river to intercept a British force led by their commander in chief and the governor-general. In a series of bitterly contested battles at Mudki (December 18), Fīroz Shāh (Fīrozpur; December 21–22), Alīwāl (January 28, 1846), and Sobrāon (February 10)—often called the First Sikh War—the Khālsā were defeated. The British annexed the territory between the Sutlej and Beās rivers; forced the Sikhs to reduce their army; and, on their failure to pay a large war indemnity, forced them to cede Jammu and Kashmir, which were then sold to Gulab Singh Dogra. A British resident was posted at Lahore to administer the rest of the Sikh kingdom during the minority of Dalip Singh.

Administrative measures taken by the resident aroused resentment among the people. The banishment of Jindan Kaur, the queen mother, on charges of conspiracy brought matters to a head in the winter of 1848 and touched off a general Sikh uprising called the Second Sikh War. A bloody but inconclusive battle was fought at Chiliānwāla (January 13, 1849); however, at Gujrāt (February 21, 1849) the Khālsā were totally defeated and laid down their arms. The Sikh kingdom was annexed, and Maharaja Dalip Singh was exiled from the Punjab.

Sikh role in the Indian Mutiny

After many years of chaos, the Punjab was administered efficiently and fairly. Consequently, when the Indian Mutiny broke out in 1857, the province stayed loyal to the British, and the Sikhs took a prominent role in suppressing the Mutiny. For this loyalty and help they were rewarded by grants of land. The proportion of Sikhs in the British Army was increased. A regulation was passed requiring Sikh soldiers to observe Khālsā traditions. With the reclamation of desert lands through an extensive system of canals, unprecedented prosperity came to the Punjab. Sikhs were the most favoured settlers. Sikh loyalty was evidenced in World War I, in which Sikhs formed more than one-fifth of the British Indian Army.

The depression that followed the war led to widespread disturbances, which climaxed on April 13, 1919, in the killing of almost 400 people at Amritsar. Sikhs also clashed with the authorities over the possession of their *gurdwārā*s (temples), which were under the control of hereditary priests. The Sikh masses turned from their British connection to join Gandhi's freedom movement. The progressive introduction of democratic reforms further reduced their earlier privileged status under British rule. Their participation on the British side in World War II was considerably less enthusiastic than it had been in 1914–18.

When the subcontinent was partitioned into India and Pakistan in 1947, the Sikh population was divided equally on both sides of the boundary line. Since the partition had been preceded by savage Sikh-Muslim riots, some 2.5 million Sikhs were compelled to leave Pakistan.

Sikhism since 1947. The government of free India abolished privileges previously extended by the British to religious minorities, including the Sikhs. Thus, the proportion of Sikhs in defense and civil services declined. The partition also adversely affected the Sikh agricultural classes, who had abandoned rich farmlands in Pakistan and changed places with Muslims of east Punjab whose holdings were much smaller. The decline in their fortunes nurtured a sense of grievance and gave birth to agitation for a Punjabi-speaking province in India in which Sikhs would form a majority of the population. This demand was conceded after the Indo-Pakistan War in 1965.

Increased wheat production during the 1970s brought unprecedented prosperity to Sikh farmers. Material improvement was accompanied by the growth of Sikh fundamentalism under the leadership of Jarnail Singh Bhindranwale. Tension increased between Sikhs and Hindus as the Akālī Religious Party (Shiromanī Akālī Dal; SAD), the predominant Sikh political party, began demanding more political and economic advantages for Sikhs.

By the early 1980s the demands of the SAD had become strongly militant, and there was an escalation in sectarian violence. The Indian government responded by arresting and imprisoning thousands of Sikhs. Armed bands, under the direction of Bhindranwale, spread a reign of terror throughout the Punjab region. Matters came to a head in 1984, when Bhindranwale and his followers entrenched themselves in the compound of the Harimandir (Golden Temple). In June the Indian Army launched an assault on the temple that killed several hundred Sikhs (including Bhindranwale) and resulted in heavy damage to the temple buildings. In October Indian Prime Minister Indira Gandhi was assassinated by two Sikh members of her bodyguard, touching off widespread Hindu violence against Sikhs. These two events caused deep resentment in the Sikh community and fueled the movement demanding the establishment of a separate Sikh state.

SIKH LITERATURE, MYTH, AND LORE

Canonical and noncanonical literature. The earliest source materials on Nānak are the *janam-sākhīs* ("life stories"), written 50 to 80 years after the death of the Gurū. Most Sikh scholars reject them and rely instead on the Gurū's compositions incorporated in the *Ādi Granth* and the *Vārs* (heroic ballads) composed by Bhāī Gurdās (died 1629). Neither Nānak's hymns nor Gurdās' *Vārs* are specific regarding the events of Nānak's life. Other historical writings date from the 18th and 19th centuries.

The canonical work

There is only one canonical work: the *Ādi Granth* ("First Book") compiled by the fifth Gurū, Arjun, in 1604. There are at least three recensions (versions) of the *Ādi Granth* that differ from each other in minor detail. The version accepted by Sikhs as authentic is said to have been revised by Gobind Singh in 1704. The *Ādi Granth* contains nearly 6,000 hymns composed by the first five Gurūs: Nānak (974), Aṅgad (62), Amar Dās (907), Rām Dās (679), and Arjun (2,218). Gobind incorporated 115 hymns written by his father, Tegh Bahādur, in it. Besides these compositions, the *Ādi Granth* contains hymns of the Bhakta saints and Muslim Ṣūfīs (notably Ravidāss, Kabīr, and Farīd Khān) and of a few of the bards attached to the courts of the Gurūs.

The *Dasam Granth* ("Tenth Book") is a compilation of writings ascribed to Gobind Singh. Scholars do not agree on the authenticity of the contents of this *Granth,* and it is not accorded the same sanctity as the *Ādi Granth*. Traditions of the Khālsā are contained in the *Rahatnāmās* (codes of conduct) by contemporaries of Gobind Singh.

Myths and lore. Although the Gurūs themselves disclaimed miraculous powers, a vast body of *sākhīs* ("stories") recounting such miracles grew up, and with them *gurdwārās* (temples) commemorating the sites where they were performed. It became an article of belief that the spirit of one Gurū passed to his successor "as one lamp lights another." This notion gained confirmation through the fact that the Gurūs used the same poetic pseudonym, "Nānak," in their compositions.

A composition about which little is known, but which has played an important role in Sikh affairs, is a collection of prophecies, *Sau Sākhī* ("Hundred Stories"), ascribed to Gobind Singh. Various versions are known to have been published prophesying changes of regimes and the advent of a redeemer who will spread Sikhism over the globe.

SIKH DOCTRINES, PRACTICES, AND INSTITUTIONS

Use of Hindu cosmogony

Doctrines. *Views on the nature of man and the universe.* Speculation on the origin of the cosmos is largely derived from Hindu texts. Sikhs accept the cyclic Hindu theory of *saṃsāra*—birth, death, and rebirth—and *karma,* whereby the nature of one's life is determined by his actions in a previous life. Humans are, therefore, equal to all other creatures, except insofar as they are sentient. Human birth is the one opportunity to escape *saṃsāra* and attain salvation.

Concept of the Khālsā. Khālsā is a concept of a "chosen" race of soldier-saints committed to a Spartan code of conduct (consisting of abstinence from liquor, tobacco, and narcotics and devotion to a life of prayer) and a crusade for *dharmayudha*—the battle for righteousness. The number five has always had mystic significance in the Punjab—"land of the five rivers." "Where there are five, there am I," wrote Gobind Singh. The first Khālsā were *pañj piyāres*—the five beloved ones. The ideal goal of all young Sikhs is to take *pahul* ("baptism") and thus become Khālsā. The *sahajdhārī* ("slow adopter") is assumed to be preparing himself gradually for the initiation.

The notion of the five Ks. The five emblems of the Khālsā, all beginning with the letter *k,* have no scriptural basis but are mentioned in the *Rahatnāmās*, written by Gobind Singh's contemporaries. The most important of the Ks is *keṣa* ("hair"), which the Khālsā must retain unshorn. A Khālsā who cuts off his hair is a *patit* ("renegade"). The sanctity of unshorn hair is older than Gurū Gobind Singh—the founder of the Khālsā—for many of the earlier Gurūs also followed the tradition (common among certain sects of Hindu ascetics as well) of letting their hair and beards grow. The other four Ks are *kaṅghā* ("comb"); *kacch* ("drawers"), worn by soldiers; *kirpān* ("sabre"); and *kārā* ("bracelet") of steel, commonly worn on the right arm.

Monotheism. Unity of the Godhead is emphasized in Sikhism. Nānak used the Hindu Vedāntic concept of *om,* the mystic syllable, as a symbol of God. To this he added the qualifications of singleness and creativity and thus constructed the symbol *ik* ("one") *om kār* ("creator"), which was later given figurative representation as ੴ. The opening lines of his morning prayer, *Japjī,* called the Mul Mantra ("Root Belief") of Sikhism, define God as the One, the Truth, the Creator, immortal and omnipresent. God is also formless (*niraṇkār*) and beyond human comprehension. Sikh scriptures use many names, both Hindu and Muslim, for God. Nānak's favourite names were Sat-Kartār ("True Creator") and Sat-Nām ("True Name"). Later the word Wāh-Gurū ("Hail Gurū") was added and is now the Sikh synonym for God.

Concepts of spiritual authority. The sole repository of spiritual authority is the *Ādi Granth.* In the event of disputes, a conclave is summoned to meet at the Akāl Takht ("Throne of the Timeless"), a building erected by the sixth Gurū, Hargobind, facing the Harimandir temple in Amritsar. Resolutions passed at the Akāl Takht have spiritual sanction. Sikh religion and politics have always been intimately connected, and belief in a Sikh state is an article of faith. "Raj karey Ga Khālsā" ("the Khālsā shall rule") is chanted at the conclusion of every service.

Aniconic orientation

Views on idolatry and rituals. Sikhism forbids representation of God in pictures and the worship of idols. Nevertheless, the *Ādi Granth* itself has become an object of intense ceremonial reverence and, as such, is known as *Granth Sahib* ("The *Granth* Personified"). *Granth Sahib* is "roused" in the morning and placed under an awning draped in fineries. Devotees do obeisance and place offerings before it. In the evening it is put to rest for the night. On festival days it is taken in procession through the streets. Most rituals centre on the *Ādi Granth.* The nonstop recitation from cover to cover by a relay of readers (*akhand-path*), which takes two days and nights, has become popular.

Social consequences of beliefs. The main consequence of Sikh belief has been a gradual breaking away from the Hindu social system and the development of Sikh separatism. The singular worship of the *Ādi Granth* excludes worship of all other objects common among Hindus (*i.e.,* the Sun, rivers, trees, etc.) and also puts a stop to the practice of ritual purifications and pilgrimages to the Ganges. Since every Sikh is entitled to read the scripture, Sikhs do not have a priestly caste similar to the Brahmans in Hinduism. Sikh insistence on commensality (eating together) at the *Gurū ka laṅgar* ("kitchen of the Gurū") destroyed the traditional Hindu pattern of caste among them and substituted a far less rigid social structure. Sikhs are grouped into three broad categories based largely on ethnic differences: Jāṭs (agricultural tribes), non-Jāṭs (erstwhile Brahmans, Kṣatriyas, and Vaiśyas—the three highest groups of the traditional Hindu social system), and Mazahabis (untouchables). The Jāṭs, though low in the caste hierarchy, are preeminent; the Mazahabis, though converts from Hindu outcastes (untouchables outside the caste sys-

Sikh social structure

tem) and still discriminated against, have a much higher status than untouchables in Hindu society. This three-tiered system is in a state of flux: among the educated urban classes it is breaking up, but in the villages a form of apartheid persists.

Practices and institutions. *The Gurū and the disciple.* The guidance of the Gurū toward the attainment of *mokṣa*—release—is absolutely essential. The Gurū or the Satgurū—true Gurū—is accorded a status only a shade below that of God. His function is to point the way to the realization of the truth, to explain the nature of reality, and to give the disciple the gift of the divine word (*nām-dān*). Although the line of Gurūs ended with Gobind Singh and Sikhs regard the *Ādi Granth* as their "living" Gurū, the practice of attaching oneself to a *sant* ("saint") and elevating him to a status of a Gurū has persisted and is widely practiced.

Recitation of nāma. Sikhism is often described as *nāmmārga* ("the way of *nāma*") because it emphasizes the constant repetition (*jap*) of the name of God and the *gurbāni* (the divine hymns of the Gurūs). *Nāma* cleanses the soul of sin and conquers the source of evil, *haumain* ("I am")—the ego. Thus tamed, the ego becomes a weapon with which one overcomes lust, anger, greed, attachment, and pride. *Nāma* stills the wandering mind and induces a superconscious stillness (*divya dṛṣṭi*); it opens the *dasam duār* ("10th gate"—the body has only nine natural orifices), through which enters divine light; and thus a person attains the state of absolute bliss.

Rites of passage and other ceremonies. No specific rites are prescribed for birth, but the practice of chanting the first five verses of Nānak's *Japjī* is observed among some Sikhs when a child is born. A few days later the child is brought to the *gurdwārā*. The *Ādi Granth* is opened and the child given a name beginning with the first letter of the first word on the left page. When a child has learned some Gurmukhi script, he is initiated into reading the *Ādi Granth*. The most important ceremony is that of *pahul* ("baptism"), usually administered at puberty. The initiate takes *amrit* ("nectar") and is admitted to the Khālsā fraternity. During a Sikh marriage ceremony (*anand karaj*), the groom and bride are required to go around the *Ādi Granth* four times to the chanting of wedding hymns. On death, there is continual chanting of hymns until the body is prepared for cremation. A final *ardāsā* ("supplicatory") prayer is said before the funeral pyre is lit. Ashes of the dead are usually deposited in a river, preferably one of the Hindu sacred rivers, such as the Ganges.

Pilgrimage places

Sacred times and places. Early hours of the dawn are ambrosial hours (*amritvelā*) most appropriate for prayer and meditation. Though not specifically prescribed as such, *gurdwārā*s with historical associations are, in fact, places of pilgrimage. Preeminent among them is the Harimandir at Amritsar, the holiest shrine of the Sikhs. Nankāna, the birthplace of Nānak (now in Pakistan), comes second. There are also five thrones (*takht*s) that are accorded special sanctity; these are at Amritsar, Anandpur, Patiāla, Patna, and Nānded. The last four are the places associated with Gobind Singh. From all of them, proclamations can be made to all the Khālsā.

The first Sikh place of worship was built by Nānak at Kartārpur and was, like Hindu temples, known as *dharamsālā* ("place of faith"). At a later stage, a Sikh temple was called a *gurdwārā*, meaning "gateway to the Gurū." There are more than 200 historical *gurdwārā*s associated with the Gurūs, which are controlled by the Shiromanī Gurdwārā Prabaṅdhak Committee (SGPC) set up by the Sikh Gurdwārās Act of 1925.

In addition to historical *gurdwārā*s, every place with a sizable Sikh population is likely to have a *gurdwārā* of its own. In well-to-do homes, a room is often set apart for this purpose. The only object of worship is the *Ādi Granth*. Sikhs observe all festivals celebrated by the Hindus of northern India. In addition, they celebrate the birthdays of the first and the last Gurūs and the martyrdom of the fifth (Arjun) and the ninth (Tegh Bahādur). The biggest fair is on the first of Baisākh (mid-April), which is also the birthday of the Khālsā itself.

The Khālsā Saṅgat. The Saṅgat ("Congregation") is usually called the Sādh-saṅgat ("Congregation of Holy Men") and thus is invested with sanctity. The Saṅgat in each *gurdwārā* elects its own governing body, and decisions are taken by vote. As a rule, women do not participate in the deliberations. The SGPC at Amritsar is the general governing body of Sikhism.

Minor subgroups and orders

Sectarian differences. The first religious order in Sikhism, the Udāsīs, emerged as a result of a disputed succession. The order was formed by Nānak's elder son, Śri Chand, who was not chosen to succeed his father as Gurū. The order, which reveres both Nānak and his son, inclined toward asceticism and later furnished priests (*mahants*) for *gurdwārā*s. They were ousted from control by the SGPC in 1925. Followers of Rām Rāī, who was passed over by his father, Har Rāī (seventh Gurū), in favour of a younger son, Hari Krishen (eighth Gurū), broke away to become Rām Rāiyās. They have their headquarters in Dehra Dūn, Uttar Pradesh.

Khālsā who do not believe that the line of Gurūs ended with Gobind Singh have continued the tradition of having a living Gurū. Among these, the Bandaī Khālsā (followers of Bandā Bahādur) are now extinct, but the Nāmdhārīs and Nirañkārīs worship living Gurūs.

Sikh welfare and educational institutions. The SGPC is the chief welfare organization of the Sikhs. The Sikh Educational Conferences, which have been meeting annually since 1908, are the chief educational organizations and are credited with the establishment of a large number of schools. In 1965 two socioreligious organizations, the Gurū Gobind Singh Foundation and the Gurū Nānak Foundation, endowed many university chairs for the study of Sikhism and for the publication of material on Sikh history and religion.

BIBLIOGRAPHY. General works are W. OWEN COLE and PIARA SINGH SAMBHI, *The Sikhs: Their Religious Beliefs and Practices* (1978, reprinted 1985), and *A Popular Dictionary of Sikhism* (1990); and HARBANS SINGH (ed.), *The Encyclopaedia of Sikhism* (1992–). JOSEPH D. CUNNINGHAM, *A History of the Sikhs,* new and rev. ed. (1918, reissued 1972), was the first scholarly work on the Sikhs up to the first Anglo-Sikh war of 1845–46. His interpretation of Sikhism as an eclectic Hindu-Muslim faith remained unquestioned until McLeod's work in 1968. W.H. MCLEOD, *Gurū Nānak and the Sikh Religion* (1968, reissued 1976), casts serious doubt on source material on the life of Nānak, rejects the theory of Sikhism as an eclectic faith, and asserts that it is a branch of Hindu Vaiṣṇavism tinged with yogism. McLeod's *Evolution of the Sikh Community* (1975) is an excellent history and analysis, and *The Sikhs: History, Religion, and Society* (1989), originally presented as lectures, is a further examination of Sikhism. Other studies include SHER SINGH, *Philosophy of Sikhism* (1944), a scholarly work interpreting Sikhism as an offshoot of Vaiṣṇavite Hinduism; and KHUSHWANT SINGH, *Ranjit Singh, Maharajah of the Punjab* (1962, reprinted 1985), a study based on Persian, Punjabi, and English sources, and *A History of the Sikhs,* 2 vol. (1963–66, reissued 1984), which interprets Sikhism as an aspect of Punjabi nationalism.

Specific aspects of Sikhism are treated in W. OWEN COLE, *Sikhism and Its Indian Context: 1469–1708* (1984), exploring the relations between early Sikhism and other Indian religious beliefs and practices; INDIA, *White Paper on the Punjab Agitation* (1984), dealing with the storming of the Golden Temple by the Indian Army; RAJIV A. KAPUR, *Sikh Separatism: The Politics of Faith* (1986), assessing the Sikh separatist movement since the 1870s; N. GERALD BARRIER and VERNE A. DUSENBERY (eds.), *The Sikh Diaspora: Migration and the Experience Beyond Punjab* (1989), studies of Sikhs leaving the Punjab and of their communities abroad; and W.H. MCLEOD, *Who Is a Sikh? The Problem of Sikh Identity* (1989), an important examination of the issue of Sikh identity.

A selection of sacred hymns from the Sikh scriptures may be found in *Selections from the Sacred Writings of the Sikhs,* trans. by TRILOCHAN SINGH *et al.* (1960, reissued 1973). MAX ARTHUR MACAULIFFE, *The Sikh Religion,* 6 vol. in 3 (1909, reprinted 1990), is a compilation of all the legends about the Sikh Gurūs based on the *janam-sākhī*s and on saints whose writings are in the Sikh scriptures. Also of interest is W.H. MCLEOD, *Early Sikh Tradition: A Study of the Janam-sākhīs* (1980).

A bibliography of works published in English since 1965 is found in PRIYA MUHAR RAI (compiler), *Sikhism and the Sikhs: An Annotated Bibliography* (1989). W.H. MCLEOD (ed. and trans.), *Textual Sources for the Study of Sikhism* (1984, reprinted 1990), is also useful. (K.S./Ed.)

Slavery

There is no consensus on what a slave was or on how the institution of slavery should be defined. Nevertheless, there is general agreement among historians, anthropologists, economists, sociologists, and others who study slavery that most of the following characteristics should be present in order to term a person a slave. The slave was a species of property; thus, he belonged to someone else. In some societies slaves were considered movable property, in others immovable property, like real estate. They were objects of the law, not its subjects. Thus, like an ox or an ax, the slave was not ordinarily held responsible for what he did. He was not personally liable for torts or contracts. The slave usually had few rights and always fewer than his owner, but there were not many societies in which he had absolutely none. As there are limits in most societies on the extent to which animals may be abused, so there were limits in most societies on how much a slave could be abused. The slave was removed from lines of natal descent. Legally and often socially, he had no kin. No relatives could stand up for his rights or get vengeance for him. As an "outsider," "marginal individual," or "socially dead person" in the society where he was enslaved, his rights to participate in political decision making and other social activities were fewer than those enjoyed by his owner. The product of a slave's labour could be claimed by someone else, who also frequently had the right to control his physical reproduction.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 513.

This article is divided into the following sections:

Historical survey 289
Slave-owning societies
Slave societies
The international slave trade
The abolition and transformation of slavery
The law of slavery 293
Sources of slavery law
Legal definitions of slavery
Master–slave legal relationships
Legal relationships between slave owners
Legal relationships between slaves and free strangers
Laws of manumission
The sociology of slavery 296
The slave as outsider
Attitudes toward slavery: the matter of race
Slave occupations
Slave demography
Slave protest
Slave culture 299
Bibliography 299

Slavery was a form of dependent labour performed by a nonfamily member. The slave was deprived of personal liberty and the right to move about geographically as he desired. There were likely to be limits on his capacity to make choices with regard to his occupation and sexual partners as well. Slavery was usually, but not always, involuntary. If not all of these characterizations in their most restrictive forms applied to a slave, the slave regime in that place is likely to be characterized as "mild"; if almost all of them did, then it ordinarily would be characterized as "severe."

Sources for slaves

Slaves were generated in many ways. Probably the most frequent was capture in war, either by design, as a form of incentive to warriors, or as an accidental by-product, as a way of disposing of enemy troops or civilians. Others were kidnapped on slave-raiding or piracy expeditions. Many slaves were the offspring of slaves. Some people were enslaved as a punishment for crime or debt, others were sold into slavery by their parents, other relatives, or even spouses, sometimes to satisfy debts, sometimes to escape starvation. A variant on the selling of children was the exposure, either real or fictitious, of unwanted children, who were then rescued by others and made slaves. Another source of slavery was self-sale, undertaken sometimes to obtain an elite position, sometimes to escape destitution.

Slavery existed in a large number of past societies whose general characteristics are well-known. It was rare among primitive peoples, such as the hunter-gatherer societies, because for slavery to flourish, social differentiation or stratification was essential. Also essential was an economic surplus, for slaves were often consumption goods who themselves had to be maintained rather than productive assets who generated income for their owner. Surplus was also essential in slave systems where the owners expected economic gain from slave ownership.

Ordinarily there had to be a perceived labour shortage, for otherwise it is unlikely that most people would bother to acquire or to keep slaves. Free land and, more generally, open resources were often a prerequisite for slavery; in most cases where there were no open resources, nonslaves could be found who would fulfill the same social functions at lower cost. Last, some centralized governmental institutions willing to enforce slave laws had to exist, or else the property aspects of slavery were likely to be chimerical. Most of these conditions had to be present in order for slavery to exist in a society; if they all were, until the abolition movement of the 19th century swept throughout most of the world, it was almost certain that slavery would be present. Although slavery existed almost everywhere, it seems to have been especially important in the development of two of the world's major civilizations, Western (including ancient Greece and Rome) and Islāmic.

Two types of slavery

There have been two basic types of slavery throughout recorded history. The most common has been what is called household, patriarchal, or domestic slavery. Although domestic slaves occasionally worked outside the household, for example, in haying or harvesting, their primary function was that of menials who served their owners in their homes or wherever else the owners might be, such as in military service. Slaves often were a consumption-oriented status symbol for their owners, who in many societies spent much of their surplus on slaves. Household slaves sometimes merged in varying degrees with the families of their owners, so that boys became adopted sons or women became concubines or wives who gave birth to heirs. Temple slavery, state slavery, and military slavery were relatively rare and distinct from domestic slavery, but in a very broad outline they can be categorized as the household slaves of a temple or the state.

The other major type of slavery was productive slavery. It was relatively infrequent and occurred primarily in classical Athenian Greece and Rome and in the post-Columbian circum-Caribbean New World. It also was found in 9th-century Iraq, among the Kwakiutl Indians of the American Northwest, and in a few areas of sub-Saharan Africa in the 19th century. Although slaves also were employed in the household, slavery in all of those societies seems to have existed predominantly to produce marketable commodities in mines or on plantations.

A major theoretical issue is the relationship between productive slavery and the status of a society as a slave or a slave-owning society. In a slave society, slaves composed a significant portion (at least 20–30 percent) of

the total population, and much of that society's energies were mobilized toward getting and keeping slaves. In addition the institution of slavery had a significant impact on the society's institutions, such as the family, and on its social thought, law, and economy. It seems clear that it was quite possible for a slave society to exist without productive slavery; the known historical examples were concentrated in Africa and Asia. It is also clear that most of the slave societies have been concentrated in Western (including Greece and Rome) and Islāmic civilizations. In a slave-owning society slaves were present, but in smaller numbers, and they were much less the focus of the society's energies.

Slavery was a species of dependent labour differentiated from other forms primarily by the fact that in any society it was the most degrading and most severe. Slavery was the prototype of a relationship defined by domination and power. But throughout the centuries man has invented other forms of dependent labour besides slavery, including serfdom, indentured labour, and peonage. The term serfdom is much overused, often where it is not appropriate (always as an appellation of opprobrium). In the past a serf usually was an agriculturalist, whereas, depending upon the society, a slave could be employed in almost any occupation. Canonically, serfdom was the dependent condition of much of the western and central European peasantry from the time of the decline of the Roman Empire until the era of the French Revolution. This included a "second enserfment" that swept over central and some of eastern Europe in the 15th and 16th centuries. Russia did not know the "first enserfment"; serfdom began there gradually in the mid-15th century, was completed by 1649, and lasted until 1906. Whether the term serfdom appropriately describes the condition of the peasantry in other contexts is a matter of vigorous contention. Be that as it may, the serf was also distinguished from the slave by the fact that he was usually the subject of the law—*i.e.*, he had some rights, whereas the slave, the object of the law, had significantly fewer rights. The serf, moreover, was usually bound to the land (the most significant exception was the Russian serf between about 1700 and 1861), whereas the slave was always bound to his owner; *i.e.*, he had to live where his owner told him to, and he often could be sold by his owner at any time. The serf usually owned his means of production (grain, livestock, implements) except the land, whereas the slave owned nothing, often not even the clothes on his back. The serf's right to marry off his lord's estate often was restricted, but the master's interference in his reproductive and family life ordinarily was much less than was the case for the slave. Serfs could be called upon by the state to pay taxes, to perform corvée labour on roads, and to serve in the army, but slaves usually were exempt from all of those obligations.

Distinction between slavery and serfdom

A person became an indentured servant by borrowing money and then voluntarily agreeing to work off the debt during a specified term. In some societies indentured servants probably differed little from debt slaves (*i.e.*, persons who initially were unable to pay off obligations and thus were forced to work them off at an amount per year specified by law). Debt slaves, however, were regarded as criminals (essentially thieves) and thus liable to harsher treatment. Perhaps as many as half of all the white settlers in North America were indentured servants, who agreed to work for someone (the purchaser of the indenture) upon arrival to pay for their passage. Some indentured servants alleged that they were treated worse than slaves; the economic logic of the situation was that slave owners thought of their slaves as a long-term investment whose value would drop if maltreated, whereas the short-term (typically four years) indentured servants could be abused almost to death because their masters had only a brief interest in them. Practices varied, but indenture contracts sometimes specified that the servants were to be set free with a sum of money, sometimes a plot of land, perhaps even a spouse, whereas for manumitted slaves the terms usually depended more on the generosity of the owner.

Peons

Peons were either persons forced to work off debts or criminals. Peons, who were the Latin-American variant of debt slaves, were forced to work for their creditors to pay off what they owed. They tended to merge with felons because people in both categories were considered criminals, and that was especially true in societies where money fines were the main sanction and form of restitution for crimes. Thus, the felon who could not pay his fine was an insolvent debtor. The debt peon had to work for his creditor, and the labour of the criminal peon was sold by the state to a third party. Peons had even less recourse to the law for bad treatment than did indentured servants, and the terms of manumission for the former typically were less favourable than for the latter.

HISTORICAL SURVEY

The origins of slavery are lost to human memory. It is sometimes hypothesized that at some moment it was decided that persons detained for a crime or as a result of warfare would be more useful if put to work in some way rather than if killed outright and discarded or eaten. But both if and when that first occurred is unknown.

Slave-owning societies. Slavery is known to have existed as early as the Shang dynasty (18th–12th century BC) in China. It has been studied thoroughly in ancient Han China (206 BC–AD 25), where perhaps 5 percent of the population was enslaved. Slavery continued to be a feature of Chinese society down to the 20th century. For most of that period it appears that slaves were generated in the same ways they were elsewhere, including capture in war, slave raiding, and the sale of insolvent debtors. In addition, the Chinese practiced self-sale into slavery, the sale of women and children (to satisfy debts or because the seller could not feed them), and the sale of the relatives of executed criminals. Finally, kidnapping seems to have produced a regular flow of slaves at some times. The go-between or middleman was an important figure in the sale of local people into slavery; he provided the distance that made such slaves into outsiders, for the purchasers did not know their origins. Chinese family boundaries were relatively permeable, and some owners established kinlike relations with their slaves; male slaves were appointed as heirs when no natural offspring existed. As was also the case in other slave-owning societies, slaves in China were often luxury consumption items who constituted a drain on the economy. The reasons China never developed into a slave society are many and complex, but certainly an abundance of non-slave labour at low prices was one of the major ones.

Slavery in China

Korea had a very large slave population, ranging from a third to half of the entire population for most of the millennium between the Silla period and the mid-18th century. Most of the Korean slaves were indigenously generated. In spite of their numbers, slaves seem to have had little impact on other institutions, and thus the society can be categorized as a slave-owning one.

Slavery existed in ancient India, where it is recorded in the Sanskrit Laws of Manu of the 1st century BC. The institution was little documented until the British colonials in the 19th century made it an object of study because of their desire to abolish it. In 1841 there were an estimated 8,000,000 or 9,000,000 slaves in India, many of whom were agrestic or predial slaves, that is, slaves who were attached to the land they worked on but who nevertheless could be alienated from it. Malabar had the largest proportion of slaves, about 15 percent of the total population. The agrestic slaves initially were subjugated communities. The remainder of the slaves was recruited individually by purchase from dealers or parents or by self-sale of the starving, and they can be classified as household slaves. Slavery in Hindu India was complicated by the slave owners' ritual need to know the origins of their slaves, which explains why most of them were of indigenous origin. Although there were exceptions, slaves were owned primarily for prestige.

Slavery was widely practiced in other areas of Asia as well. A quarter to a third of the population of some areas of Thailand and Burma were slaves in the 17th through the 19th centuries and in the late 19th and early 20th centuries, respectively. But not enough is known about them to say that they definitely were slave societies.

Other societies in the Philippines, Nepal, Malaya, In-

donesia, and Japan are known to have had slavery from ancient until fairly recent times. The same was true among the various peoples inhabiting the regions of Central Asia: the peoples of Sogdiana, Khorezm, and other advanced civilizations; the Mongols, the Kalmyks, the Kazakhs; and the numerous Turkic peoples, most of whom converted to Islām.

Slave-owning societies in the New World

In the New World some of the best-documented slave-owning societies were the Klamath and Pawnee and the fishing societies, such as the Yurok, that lived along the coast from what is now Alaska to California. Life was easy in many of those societies, and slaves are known to have sometimes been consumption goods that were simply killed in potlatches.

Other Amerindians, such as the Creek of Georgia, the Comanche of Texas, the Callinago of Dominica, the Tupinambá of Brazil, the Inca of the Andes, and the Tehuelche of Patagonia, also owned slaves. Among the Aztecs of Mexico, slavery generally seems to have been relatively mild. People got into the institution through self-sale and capture and could buy their way out relatively easily. Slaves were often used as porters in the absence of draft animals in Mesoamerica. The fate of other slaves was less pleasant: chattels purchased from the Mayans and others were sacrificed in massive numbers. Some of the sacrifices may have been eaten by the social elite.

In England about 10 percent of the population entered in the Domesday Book in 1086 were slaves, with the proportion reaching as much as 20 percent in some places. Slaves were also prominent in Scandinavia during the Viking era, AD 800–1050, when slaves for use at home and for sale in the international slave markets were a major object of raids. Slaves also were present in significant numbers in Scandinavia both before and after the Viking era.

Continental Europe—France, Germany, Poland, Lithuania, and Russia—all knew slavery. Russia was essentially founded as a by-product of slave raiding by the Vikings passing from Scandinavia to Byzantium in the 9th century, and slavery remained a major institution there until the early 1720s, when the state converted the household slaves into house serfs in order to put them on the tax rolls. House serfs were freed from their lords by an edict of Tsar Alexander II in 1861. Many scholars argue that the Soviets reinstituted a form of state slavery in the Gulag camps that flourished until 1956.

Slavery was much in evidence in the Middle East from the beginning of recorded history. It was treated as a prominent institution in the Babylonian Code of Hammurabi of *c.* 1750 BC. Slaves were present in ancient Egypt and are known to have been murdered to accompany their deceased owners into the afterlife. It once was believed that slaves built the great pyramids, but contemporary scholarly opinion is that the pyramids were constructed by peasants when they were not occupied by agriculture. Slaves also are mentioned prominently in the Bible among the Hebrews in Palestine and their neighbours.

Slaves were owned in all Islāmic societies, both sedentary and nomadic, ranging from Arabia in the centre to North Africa in the west and to what is now Pakistan and Indonesia in the east. Some Islāmic states, such as the Ottoman Empire, the Crimean Khanate, and the Sokoto caliphate, must be termed slave societies because slaves there were very important numerically as well as a focus of the polities' energies.

Slaves have been owned in black Africa throughout recorded history. In many areas there were large-scale slave societies, while in others there were slave-owning societies. Slavery was practiced everywhere even before the rise of Islām, and black slaves exported from Africa were widely traded throughout the Islāmic world. Approximately 18,000,000 Africans were delivered into the Islāmic trans-Saharan and Indian Ocean slave trades between 650 and 1905. In the second half of the 15th century Europeans began to trade along the west coast of Africa, and by 1867 between 7,000,000 and 10,000,000 Africans had been shipped as slaves to the New World. Although some areas of Africa were depleted by slave raiding, on balance the African population grew after the establishment of the transatlantic slave trade because of new food crops introduced from the New World, particularly manioc, corn (maize), and possibly peanuts (groundnuts). The relationship between African and New World slavery was highly complementary. African slave owners demanded primarily women and children for labour and lineage incorporation and tended to kill males because they were troublesome and likely to flee. The transatlantic trade, on the other hand, demanded primarily adult males for labour and thus saved from certain death many adult males who otherwise would have been slaughtered outright by their African captors. After the end of the transatlantic trade, a few African societies at the end of the 19th century put captured males to productive work as slaves, but this usually was not the case before that time.

Slavery in black Africa

Slave societies. The first known major slave society was that of Athens. In the early Archaic period the elite worked its estates with the labour of fellow citizens in bondage (often for debt). After the lawgiver Solon abolished citizen slavery about 594 BC, wealthy Athenians came to rely on enslaved peoples from outside Attica. The prolonged wars with the Persians and other peoples provided many slaves, but the majority of slaves were acquired through regular trade with non-Greek peoples around the Aegean. At the time of classical Athens (the 5th through the 3rd centuries BC) slaves constituted about a third of the population. A particularly noteworthy locus of slave employment was the Laurium silver mines, where private individuals could pick out a lode and put their slaves to mining it. As in all other slave societies, it was the profitability of slavery that determined its preeminence in Athens. (Also important were political conditions that made the gross exploitation of citizens impossible.) Slaves were responsible for the prosperity of Athens and the leisure of the aristocrats, who had time to create the high culture now considered the beginning of Western civilization. The existence of large-scale slavery was also responsible, it seems logical to believe, for the Athenians' thoughts on freedom that are considered a central part of the Western heritage. Athenian slave society was finally destroyed by Philip II of Macedonia at the battle of Chaeronea (338 BC), when, on the motion of Lycurgus, many (but not all) slaves were freed.

Athenian slave society

The next major slave society was Roman Italy between about the 2nd century BC and the 4th century AD. Initially, Rome was a polity consisting primarily of small farmers. But the process of creating the empire took them away from their farms for extended periods, and the prolonged wars of conquest in Spain and the eastern Mediterranean during the 3rd and 2nd centuries BC created a great flood of captives. Nothing was more logical than to put the captives to work farming, especially the olives and grapes that created much of the prosperity of the late republic and the principate. Slaves and freedmen were responsible for much of the empire's commodity production, and in the early principate they ran its governmental bureaus as well. The conditions were right to put the captives to work: private ownership of land; developed commodity production and markets; a perceived shortage of internal labour supply; and an appropriate moral, political, and legal climate. Roughly 30 percent of the population was enslaved. Roman slave society ended as the slaves were legally converted into coloni, or serfs, and the lands became populated and the frontiers so remote that finding great numbers of outsider slaves was increasingly difficult.

Roman slave society

Some lesser Islāmic slave societies are also of interest. One is the Baghdad caliphate founded in the 7th and lasting through the 10th century. Many tens of thousands of military captives were imported from Sogdiana, Khazaria, and other Central Asian locales. In the 9th and 10th centuries several tens of thousands of black Zanj slaves were imported from Zanzibar to Lower Iraq, where they constituted more than half the total population and were put to work to clear saline lands for irrigation and to cultivate sugar. More long-term was the slavery practiced in the Crimean Khanate between roughly 1475 and its liquidation by the Russian empress Catherine the Great in 1783. The Crimean Tatar society was based on raiding the neighbouring Slavic and Caucasian sedentary societies and selling the captives into the slave markets of Eurasia. Approximately 75 percent of the Crimean population

consisted of slaves or freedmen, and much of the free population was highly predatory, engaged either in the gathering of slaves or in the selling of them. It is known that for every slave the Crimeans sold in the market, they killed outright several other people during their raids, and a couple more died on the way to the slave market. The reasons for the transition of the Crimean Khanate from a slave-owning society to a slave society have not been studied in detail. Probable reasons, however, include the combination of high demand for slaves throughout the Islāmic world, the defenselessness of the sedentary agricultural Slavs and others, and the existence of a relatively poor class of Crimean horsemen, who were led by a predatory elite that got rich by slave raiding. Crimean Tatar slave raids into Muscovy were greatly curtailed by the building of a series of walls along the frontier in the years 1636–53 and ultimately by the liquidation of the khanate in 1783.

Slavery in the Ottoman Empire

It is probable that the Ottoman Empire, and especially its centre in Turkey, should be termed a slave society. Slaves from both the white Slavic north and the black African south flowed into Turkish cities for half a millennium after the Turks seized control of much of the Balkans in the 14th century. The proportion of the population that was slave ranged from about one-fifth in Istanbul, the capital, to much less in remoter provincial areas. Perhaps only people such as the slave owners of the circum-Caribbean sugar islands and the American South were as preoccupied with slaves as were the Ottomans.

Slaves in the Ottoman Empire served in various capacities. They were janissary soldiers (see below), and they ran the empire, manned its ships, generated much of its handicraft product, and served as domestic servants and in harems. Contemporaries believed that the absolute power of the ruler was based on his military and administrative slaves. The Tanzimat enlightenment movement of the mid-19th century initiated the abolition of slavery; by the 1890s only a few slaves were being smuggled illegally into the empire, and the slave population was greatly reduced.

Other prominent Islāmic slave societies were on the east coast of Africa in the 19th century. The Arab-Swahili slave systems have been well-studied, and it is known that, depending on the date, 65 to 90 percent of the population of Zanzibar was enslaved. Close to 90 percent of the population on the Kenya coast was also enslaved, and in Madagascar half the population was enslaved. It may be assumed that similar situations prevailed elsewhere in the vicinity and also earlier, but studies to verify the proposition have not been undertaken.

Another notable Islāmic slave society was that of the Sokoto caliphate formed by Hausas in sub-Saharan Africa (northern Nigeria and Cameroon) in the 19th century. At least half the population was enslaved. That was only the most notable of the Fulani jihad states of the western and central Sudan, where between 1750 and 1900 from one- to two-thirds of the entire population consisted of slaves. In Islāmic Ghana, between 1076 and 1600, about a third of the population were slaves. The same was true among other early states of the western Sudan, including Mali (1200–1500), Segou (1720–1861), and Songhai (1464–1720). It should be noted that slavery was prominent in Ghana and Mali, and presumably elsewhere in Africa in areas for which information is not available, long before the beginnings of the transatlantic slave trade. The population of the notorious slave-trading state of the central Sudan, Ouidah (Whydah), was half-slave in the 19th century. It was about a third in Kanem (1600–1800) and perhaps 40 percent in Bornu (1580–1890). Most slaves probably were acquired by raiding neighbouring peoples, but others entered slavery because of criminal convictions or defaulting on debts (often not their own); subsequently, many of those people were sold into the international slave trade. After the limiting and then abolition of the transatlantic slave trade, a number of these African societies put slaves to work in activities such as mining gold and raising peanuts, coconuts (palm oil), sesame, and millet for the market.

Among some of the various Islāmic Berber Tuareg peoples of the Sahara and Sahel, slavery persisted at least until 1975. The proportions of slaves ranged from around 15 percent among the Adrar to perhaps 75 percent among the Gurma. In Senegambia, between 1300 and 1900, about a third of the population consisted of slaves. In Sierra Leone in the 19th century close to half the population was enslaved. In the Vai Paramount chiefdoms in the 19th century as much as three-quarters of the population consisted of slaves. Among the Ashanti and Yoruba a third were enslaved. In the 19th century over half the population consisted of slaves among the Duala of the Cameroon, the Ibo and other peoples of the lower Niger, the Kongo, and the Kasanje kingdom and Chokwe of Angola.

Caribbean slave societies

The best-known slave societies were those of the circum-Caribbean world. Slave imports to the islands of the Caribbean began in the early 16th century. Initially the islands often were settled as well by numerous indentured labourers and other Europeans, but following the triumph after 1645 of the sugar revolution (initially undertaken because superior Virginia tobacco had left the Barbadian planters with nothing to sell) and after the nature of the disease climate became known to Europeans, they came to be inhabited almost exclusively by imported African slaves. In time the estate owners moved to England, and the sugar plantations were managed by sometimes unstable and unsavoury Europeans who, with the aid of black overseers and drivers, controlled masses of slaves. About two-thirds of all slaves shipped across the Atlantic ended up in sugar colonies. By 1680 in Barbados the average plantation had about 60 slaves, and in Jamaica in 1832 about 150. The sugar plantations were among the contemporary world's largest and most profitable enterprises, paying about 10 percent on invested capital and on some occasions, such as in Barbados in the 1650s, as much as 40 to 50 percent. The proportions of slaves on the islands ranged from more than a third in Cuba, which went into the sugar and gang-labour business on a large scale only after the local planters had gained control in 1789, to 90 percent and more on Jamaica in 1730, Antigua in 1775, and Grenada up to 1834.

Slaves were of varying importance in Mesoamerica and on the South American continent. Initially slaves were imported because of a labour shortage, aggravated by the high death rate of the indigenous population after the introduction of European diseases in the early 16th century. They were brought in at first to mine gold, and they were shifted to silver mining or simply let go when gold was exhausted in the mid-16th century. In Brazil, where sugar had been tried even before its planting in the Caribbean, the coffee bush was imported from Arabia or Ethiopia via Indonesia, and it had an impact similar to that of sugar in the Caribbean. Around 1800 about half the population of Brazil consisted of slaves, but that percentage declined to about 33 percent in 1850 and to 15 percent after the shutting off of imports around 1850 combined with free immigration to raise the proportion of Europeans. In some parts of Brazil, such as Pernambuco, some two-thirds of the population consisted of Africans and their offspring.

The final circum-Caribbean slave society was what became the southern United States. Slaves first were brought to Virginia in 1619. Subsequently, Africans were transshipped to North America from the Caribbean in increasing numbers. Initially, however, the English relied for their dependent labour primarily on indentured servants from the mother country. But in the two decades of the 1660s and 1670s the laws of slave ownership were clarified (for example, Africans who converted to Christianity did no longer have to be manumitted), and the price of servants may have increased because of rising wage rates in prospering England; soon thereafter African slaves replaced English indentured labourers. Tobacco initially was the profitable crop that occupied most slaves in the Chesapeake. The invention of the cotton gin by Eli Whitney in 1793 changed the situation, and thereafter cotton culture created a huge demand for slaves, especially after the opening of the New South (Alabama, Mississippi, Louisiana, and Texas). By 1850 nearly two-thirds of the plantation slaves were engaged in the production of cotton. Cotton could be grown profitably on smaller plots than could sugar, with the result that in 1860 the average cotton plantation had only about 35 slaves, not all of

Cotton culture and slavery

whom produced cotton. During the reign of "King Cotton," about 40 percent of the Southern population consisted of black slaves; the percentage of slaves rose as high as 64 percent in South Carolina in 1720 and 55 percent in Mississippi in 1810 and 1860. More than 36 percent of all the New World slaves in 1825 were in the southern United States. Like Rome and the Sokoto caliphate, the South was totally transformed by the presence of slavery. Slavery generated profits comparable to those from other investments and was only ended as a consequence of the War Between the States.

The international slave trade. Organized commerce began in the Neolithic Period, and it may be assumed that slaves were not far behind high-value items such as amber and salt in becoming commodities. Even among relatively simple peoples one can trace the international slave trade. Thus such a trade was going on among the peoples of Siberia before the arrival of the Russians in the 16th and 17th centuries. The slaves so traded were neighbouring people captured in warfare, who were then shipped to distant points where they would be without kin and whence they would be unlikely to flee. Similar commerce in slaves occurred on nearly all continents and provided the bulk of household slaves throughout the world.

The international slave trades that provided much of the chattel for the slave societies flowed out of the great "population reservoirs." Two such reservoirs were the Slavs and contiguous agriculturalist Iranians from antiquity to the 19th century and the sub-Saharan Africans from around the beginning of the Christian Era to the middle of the 20th century. A third such reservoir probably was the Germanic, Celtic, and Romance peoples who lived north of the Roman Republic and Empire and who half a millennium later became the victims of the Vikings' slave raids. The dynamics of these raids were as follows: A large demand for slave labour prompted neighbouring peoples (typically migratory or nomadic in habit) to prey on the sedentary agriculturalists living in the reservoir. The raiders developed techniques, of which surprise was perhaps the major one, that put the settled peoples at a disadvantage, for they never knew when and where the raiders might strike. Populations in the reservoir could be completely depleted, as happened to the East Slavs living in the steppe south of the Oka and between the Volga and the Dnepr rivers from 1240 to the 1590s, or they could migrate half a continent away to escape the slave raiders, as did the Ndembu in Africa. Ruthenians, frontier Poles, Caucasians, and numerous African peoples were sorely depleted by slave raids. One alternative was to fight back, as did the Muscovite Russians and the Baya of Adamawa (now northern Cameroon in West Africa), and the consequence in both instances was the creation of an authoritarian garrison state.

Slave raiding

The international slave trades developed into elaborate networks. For example, in the 9th and 10th centuries Vikings and Russian merchants took East Slavic slaves into the Baltic. They were then gathered in Denmark for further transshipment and sold to Jewish and Arabic slave traders, who took them to Verdun and León. There some of the males were castrated. From those places the slaves were sold to harems throughout Moorish Spain and North Africa. In the 9th century the Baghdad caliphate got slaves from western Europe via Marseille, Venice, and Prague; Slavic and Turkic slaves from eastern Europe and Central Asia via Derbent, Itil, Khorezm, and Samarkand; and African slaves via Mombasa, Zanzibar, the Sudan, and the Sahara. The Mongols in the 13th century brought their slaves first to Karakorum, whence they were sold throughout Asia, and then later to Sarai on the Lower Volga, whence they were retailed throughout much of Eurasia. Following the breakup of the Golden Horde, the Crimean Tatars took their chattel to Kefe (Feodosiya) in the Crimea, whence it was transported across the Black Sea and sold throughout the Ottoman Empire and elsewhere. Arabs developed similar supply networks out of black Africa across the Sahara, across the Red Sea (from Ethiopia and Somalia), and out of East Africa, which supplied the Islāmic world and the Indian Ocean region with human chattel.

Beginning around 1500, a similar process occurred along the coast of West Africa to supply the transatlantic slave trade. The Africans were captured by other Africans in raids and then transported to the coast; one may assume that the number of casualties of African slave raiding was nearly as high as that of Crimean Tatar slave raiding. The captives, primarily adult males, were assembled on the coast by African rulers and kept in holding pens until wholesaled to European ship captains who sailed up and down the coast looking for slave cargo. (As stated above, the women and children often were not sent to the coast for export but were kept by the Africans themselves, often for incorporation into their lineages.) African rulers, who did not allow the Europeans to move inland, often conducted their wholesale business on the coast, such as at Ouidah in Dahomey (now Benin). (Because of the disease climate the Europeans also were reluctant, even unable, to move inland until the mid-19th century.) But African rulers did everything they could to encourage the European sea captains to come to their port.

The transatlantic slave trade

Once a ship was loaded, the trip, known as "the Middle Passage," usually to Brazil or an island in the Caribbean, was a matter of a few weeks to several months. Between 1500 and the end of the 19th century the time of the voyage diminished considerably. That change was important, because death rates, which ranged from around 10 to more than 20 percent on the Middle Passage, were directly proportional to the length of the voyage. The ship captains had every interest in the health of their cargo, for they were paid only for slaves delivered alive. The death rates among the European captains and crew engaged in the slave trade were at least as high as those among their cargo on the Middle Passage. Of the slave-ship crews that embarked from Liverpool in 1787, less than half returned alive.

Arriving in Brazil or the Caribbean islands, the slaves were sold at auction. The slave auctions were elaborate markets in which the prices of the slaves were determined. The auctions told the captains and their superiors what kind of cargo was in demand, usually adult males. Credit almost always was part of the transaction, and inability to collect was one of the major reasons companies went bankrupt. After the auction the slave was delivered to the new owner, who then put him to work. That also began the period of "seasoning" for the slave, the period of about a year or so when he either succumbed to the disease environment of the New World or survived it. Many slaves landed on the North American mainland before the early 18th century had already survived the seasoning process in the Caribbean.

It can be assumed that the other international slave trades were comparable in many respects to the transatlantic one, but they have not been adequately studied.

The abolition and transformation of slavery. Throughout history there have been people who in one way or another believed that slavery was not a good or natural condition. Jean Bodin (1530–96), the French founder of antislavery thought, for example, condemned the institution as immoral and counterproductive and advocated that no group of men should be excluded from the body politic. Nevertheless, remarkably few people found the institution of slavery to be unnatural or immoral until the second half of the 18th century. Even more disturbing is the persistence of slavery into the 21st century despite religious, legal, and moral sanctions against it.

Attitudes toward slavery

Slavery ended in many places without much fanfare. In most societies, such as ancient Babylonia, Israel, Egypt, or Athens, the institution of slavery had little or no connection with the society's rise or demise. In Rome slavery began to yield to tenancy and the antecedents of serfdom before the fall of the Western Empire, as the diminishing supply of slaves and the rise of their price coincided with a decline in the number of plantations in southern Italy, although debt and domestic slavery nonetheless continued to exist. In the Eastern Roman Empire (Byzantium) serfdom was the predominant form of dependent labour, and slavery was definitely secondary. Manumitting slaves became much easier, according to the laws, and the Ecloga and the Procheiron Nomos (see below) prescribed that the slaves of persons who died without testament had to be freed.

Throughout most of Europe, household slavery persisted into the early Middle Ages but was ultimately replaced by serfdom. By the end of the Middle Ages, slavery no longer existed in England, and the famous Cartwright decision during the reign of Elizabeth I (1569) held that "England was too pure an air for slaves to breathe in."

Enserfment in eastern Europe

Slavery persisted longer in eastern Europe. In Poland it was replaced by the second enserfment; the sale and purchase of slaves were forbidden in the 15th century. A similar process occurred in Lithuania, where slavery was formally abolished in 1588. In Russia it came to an end with the first enserfment: agricultural slaves were formally converted into serfs in 1679, and household slaves were converted into house serfs in 1723. In the Caucasus and in Central Asia slavery persisted until the second half of the 19th century.

The reexportation of slaves from England was challenged by a group of humanitarians led by Granville Sharpe. Chief Justice Mansfield ruled in 1772 that James Somerset, a fugitive slave from Virginia, could not be forcibly returned to the colonies by his master.

The fate of slavery in most of the rest of the world depended on the British abolition movement, which was initiated by the English Quakers in 1783 when they presented the first important antislavery petition to Parliament. They were following the Pennsylvania Quakers, who had voiced opposition to slavery in 1688. The Vermont constitution of 1777 was the first document in the United States to abolish slavery. Another sign of the spread of antislavery feeling was the declaration in the U.S. Constitution that the importation of slaves could be forbidden after 20 years (in 1808).

Abolishment of the slave trade by Britain

In 1807 the British abolished the slave trade with their colonies. In the Caribbean, slavery was abolished by British Parliamentary fiat, effective July 31, 1834, when 776,000 slaves in the British plantation colonies were freed. The British imperial emancipation can be attributed to the growing power of the philanthropic movement and a double switch in the focus of the British Empire, geographically from west (the Caribbean) to east (India) and economically from protectionism to laissez-faire.

The British move in 1807 to abolish the slave trade had an immediate impact on the juntas struggling for independence in Spanish America. The slave trade was declared illegal in Venezuela and Mexico in 1810, in Chile in 1811, and in Argentina in 1812. In 1817 Spain signed a treaty with Britain agreeing to abolish the slave trade in 1820, but the trade continued to the remaining Spanish colonies until 1880. Chile freed its black slaves in 1823; Mexico abolished slavery in 1829, and Peru in 1854.

The American antislavery movement, linked to the "Second Great Awakening," succeeded in arousing immense hostility between the non-slave-holding North, where most states had voluntarily abolished slavery by 1804, and the slaveholding South, where the "peculiar institution" became even further entrenched because of the spread of cotton cultivation. By the 1850s, however, the old abolition movement had flagged. It took political developments and forces (especially the emergence of the Free Soil movement and the conflict over the expansion of slavery), the South's secession, the Civil War, and Abraham Lincoln's Emancipation Proclamation on Jan. 1, 1863, to put slavery on the road to extinction in the United States. The proclamation was confirmed by the Thirteenth Amendment to the Constitution (1865), which put an end to slavery.

Abolishment of slavery in Brazil

Puerto Rico abolished slavery (with provisions for periods of apprenticeship) in 1873, and Cuba did so in 1880. Brazil was the last Western Hemisphere nation to abolish slavery. The British antislavery movement of the 1810s had almost put an end to the institution, but a thriving world market for coffee revitalized it in the 1820s. In 1850 Britain declared that a squadron would enter Brazilian territorial waters to seize vessels carrying slaves, and later that year Brazil responded by equating the slave trade with piracy. On May 13, 1888, all Brazilian slaves were manumitted.

The Chinese imperial government formally abolished slavery in 1906, and the law became effective on Jan. 31, 1910, when all adult slaves were converted into hired labourers and the young were freed upon reaching age 25. Slavery was legally abolished in Korea in the Gap-o reform of 1894 but remained extant in reality until 1930.

In Africa, however, the situation was much different. Slavery continued there after 1885, even though European military occupation of the continent advanced behind abolitionist rhetoric, which promoted conquest and colonial rule as a tool to end slavery. Once in control of the continent, European military and colonial authorities in Africa (most particularly sub-Saharan Africa) found themselves dependent on African leaders whose power often rested on slaves. These slaves had been acquired for sale to purchasers who were part of the Atlantic slave trade, but the British suppression of this trade and the ending of slavery throughout the Americas left Africa with a large slave population. The colonial authorities in Africa—French and German as well as British—thus declared African forms of slavery benign and tolerable.

The flows of involuntary migrants, whether nominally slave or indentured, continued into the 20th century in Southeast Asia and many other parts of the world. Female sex slavery was the root of Japanese seizures of women in occupied territories in Asia during World War II. The women were forced to provide sexual services to the occupying troops. Captive populations of Jews, Slavs, and other political prisoners were used by German industry during World War II to produce goods. The Soviet *gulag* system also produced its share of slave labour, both before and after World War II.

The enslavement of women, who were forced to provide sex for clients in various countries of the developed world, continued into the 21st century. Often these women were immigrants who were duped and forced into prostitution. Children also were enslaved for sex and labour exploitation, sold by their impoverished parents. Immigrants, brought to a developed country under false pretenses, became virtual slaves to those who had provided their passage, and they were forced to work as servants, prostitutes, or labourers (producing clothing, for example) until they were able to pay off the debt they owed.

A persistent form of slavery was debt bondage, an age-old system whereby one could be sold into slavery to pay a debt. Such slavery persisted into the 21st century in parts of South America and Asia.

Abolishment efforts in the 20th century

The abolition movement responded to these continuing and new forms of slavery by internationalizing its structure and, with publicity campaigns, embarrassing sovereign governments that shelter private slavers. The British and Foreign Anti-Slavery Society became Anti-Slavery International in 1909 and exposed the Portuguese abuses in São Tomé as well as enslavement of Indian rubber gatherers in Brazil by a British corporation. It lobbied the League of Nations (1926) to create a Standing Committee of Experts on Slavery in 1932, which surveyed colonial contract labour systems around the world for slavery-like abuses. Since World War II it has pressed the United Nations through the Office of the High Commissioner for Human Rights to remain aware of continuing practices and to work for reform worldwide.

Despite all this international attention, more recently, centring on Brazilian Indians, Myanmar, Mauritania, and The Sudan, and despite official denials by governments, antislavery groups estimated that 27 million people were enslaved at the beginning of the 21st century, more than in any previous historical period.

THE LAW OF SLAVERY

Sources of slavery law. By definition slavery must be sanctioned by the society in which it exists, and such approval is most easily expressed in written norms or laws. Thus it is not accidental that even the briefest code of a relatively uncomplicated slave-owning society was likely to contain at least a few articles on slavery.

Both slave-owning and slave societies that were part of the major cultural traditions borrowed some of their laws about slavery from the religious texts of their respective civilizations. Principles regarding slavery that proved to be either unprofitable or unworkable were among the first to be discarded. An obvious example is provided by the Old Testament law that Hebrew slaves were to be manumitted

after six years (Exodus 21:2; Deuteronomy 15:12). A similar general recommendation that slaves be freed after six years in bondage was adhered to by many Islāmic slave-owning societies; it helps to account for the ferocity and frequency of their slave raids, for they had a need for constant replenishment of their slave supplies. In Christian slave societies, on the other hand, the principle that the tenure of slavery should be limited was almost completely ignored.

Practically every society that possessed slaves wrote about them in its laws, and thus only a few codes can be mentioned here. The ancient Mesopotamian laws of Eshnunna (*c.* 1900 BC) and the Code of Hammurabi had a number of articles devoted to slavery, as did the Pentateuch. In ancient India the Laws of Manu of the 1st century BC contained numerous laws on slaves.

Roman slave law

Little is known about the Athenian law of slavery, but the Roman law of slavery was extraordinarily elaborate. Roman law was summed up in the great Pandects of Justinian of AD 533, and some of its slave norms later found their way into the Byzantine Ecloga (which incorporated Syrian norms as well) of AD 726 and, more deliberately, into the Procheiron Nomos of AD 867–879. Romano-Byzantine norms also found their way into the Bulgarian Court Law for the People ("Zakon Sudnyi Liudem") of the end of the 9th century and the 13th-century Ethiopian Fetha Nagast.

The European "barbarian" (Germanic) codes, which first appeared in the 5th century AD and remained in effect for about half a millennium, were derived from customary law influenced by Roman law. The slave statutes of the Russian Russkaya Pravda of the 11th–13th centuries were all clearly of native East Slavic origin. The same was true of the Muscovite court handbooks (Sudebniki) of 1497, 1550, 1589, and 1606. The Muscovite Russians had a special government office to deal with slavery matters, the Slavery Chancellery (1571–1704), and its practice became the basis of chapter 20 of the great Ulozhenie of 1649, which constituted 119 of the 967 articles of the code; other articles dealt with slavery as well.

Islāmic slave law

The Qur'ān was the fundamental starting point for Islāmic law (Sharī'ah), including the law of slavery. It was supplemented by the *ijmā'*, the scholarly legal consensus, and the *qiyās*, juristic reasoning by analogy. Islāmic law regulated in detail every part of the institution of slavery, from the jihad (holy war) and the distribution of booty to the treatment of slaves and emancipation. The last Islāmic slave law was promulgated in 1936 by King Ibn Sa'ūd of Saudi Arabia, which restated the teachings of the Qur'ān. It also required owners to register slaves with the government and licensed slave traders.

Some sub-Saharan African societies followed Islāmic law; others had their own. The latter ordinarily were not systematized until the European colonization movement, and so their law of slavery was oral common law.

Slavery was a relatively prominent institution in the Chinese T'ang Code of the 7th century AD. Subsequently it was mentioned in every Chinese law down to the 20th century and was also important in the Korean legal system. The slavery norms of the Mongol Great Yassa of Genghis Khan were locally generated, but subsequent Mongol law reveals considerable influence of the T'ang Code.

The circum-Caribbean world had several basic laws of slavery. The slave law of the Spanish-speaking colonies and then independent countries was based on the Siete Partidas of 1263–65 of Alfonso X of Castile and Léon and the Spanish Slave Code of 1789. Another important code in Latin America was Louis XIV's Code Noir of 1685. The Louisiana Slave Code of 1824 was based on the Siete Partidas and the Code Napoléon.

The Danish Virgin Islands had two largely locally generated codes of 1733 and 1755, although they were approved by the colonial administration of Denmark. The English colonies were completely autonomous, for England had no law of slavery from which to borrow. The first code was that of Barbados of 1688, whose origins are unknown. It was imitated by the South Carolina code of 1740. Beginning with Virginia in 1662, each colony in North America worked out its own ex post facto law of slavery before independence, a process that continued after the creation of the United States and until the Civil War. Slavery is mentioned only three times and referred to at most 10 times (and then only indirectly) in the U.S. Constitution, and, except for a handful of measures on fugitives, there was no federal slave law. The basic protection for the institution of slavery was the Tenth Amendment of 1791, the reserved powers clause, which left the issue of slavery and other matters to the states.

Legal definitions of slavery. Some of the definitions of slavery discussed above were legal, but the majority were not. This section focuses exclusively on legal definitions of slavery. Most groups, whether national or religious, forbade the enslavement of their fellows; thus, the Spanish could not enslave Spaniards, Arabs could not enslave Arabs, and Christians and Muslims could not enslave their coreligionists. Legally the slave ordinarily had to be an outsider. In law the slave was usually defined as property, and the question then was whether he was movable property (chattel) or real property. In most societies he was movable property but in some, real property.

Some societies, such as Muscovy in the 16th and 17th centuries, had different legal categories of slaves. There some slaves were inherited, others were purchased forever, others for a limited time could become perpetual slaves, and still others for specific functions such as estate managers. Different varieties or gradations of slaves were found elsewhere as well, as in China and in certain African societies.

Master–slave legal relationships. The master–slave relationship was the cornerstone of the law of slavery, and yet it was an area about which the law often said very little. In many societies the subordination of the slave to his owner was supposed to be complete; in general, the more complete an owner's control over his slave, the less the law was likely to say about it.

The owner's right to kill his slave

A major touchstone of the nature of a slave society was whether or not the owner had the right to kill his slave. In most Neolithic and Bronze Age societies slaves had no right to life, for slaves from ancient Egypt and the Eurasian steppes were buried alive or killed to accompany their deceased owners into the next world. Among the Northwest Coast Tlingit, slave owners killed their slaves in potlatches to demonstrate their contempt for property and wealth; they also killed old or unwanted slaves and threw their bodies into the Pacific Ocean. An owner could kill his slave with impunity in Homeric Greece, ancient India, the Roman Republic, Han China, Islāmic countries, Anglo-Saxon England, medieval Russia, and many parts of the American South before 1830.

The slave's right to life

That was not the case in other societies. The Hebrews, the Athenians, and the Romans under the principate restricted the right of slave owners to kill their human chattel. The Code of Justinian changed the definition of the slave from a thing to a person and prescribed the death penalty for an owner who killed his slave by torture, poison, or fire. Spanish law of the 1260s and 1270s denied owners the right to kill their slaves. Lithuanian and Muscovite law forbade the killing, maiming, or starving of a returned fugitive slave. Ch'ing Chinese law punished a master who killed his slave, and that punishment was more severe if the slave had done no wrong. The Aztecs under some circumstances put to death a slave owner who killed his slave. No society, on the other hand, had the slightest sympathy for the slave who killed his owner. Roman law even prescribed that all other slaves living under the same roof were to be put to death along with the slave who had committed the homicide.

Assault and general brutality were other concerns of the law of slavery. In antiquity slaves often had the right to take refuge in a temple to escape cruel owners, but that sometimes afforded little protection. The ancient Franks and the Germans warned owners against cruelty. The Code of Justinian and the Spanish Siete Partidas deprived cruel owners of their slaves, and that tradition went into the Louisiana Black Code of 1806, which made cruel punishment of slaves a crime. In modern societies brutality and sadistic murder of slaves by their owners were rarely condoned on the grounds that such episodes demoralized

other slaves and made them rebellious, but few slave owners were actually punished for maltreating their slaves. In the American South 10 codes prescribed forced sale to another owner or emancipation for maltreated slaves. Nevertheless, cases such as *State* v. *Hoover* (North Carolina, 1839) and *State* v. *Jones* (Alabama, 1843) were considered sensational because slave owners were punished for savagely "correcting" their slaves to death.

It was not an axiom of the master–slave relationship that the former automatically had sexual access to the latter. That was indeed the case in most societies, ranging from the ancient Middle East, Athens, and Rome to Africa, all Islāmic countries, and the American South. Places such as Muscovy, however, forbade owners to rape their female slaves, while the Chinese and the Lombards forbade the raping of married slave women. More problematic were sexual relations between mistresses and male slaves. Athens and Rome both put the slave to death, and Byzantine law prescribed that the mistress was to be executed and the slave to be burned alive. The Danish Virgin Islands' laws of 1741, 1755, and 1783, in an attempt to protect northern Europeans from African "contamination," prescribed a fine of 2,000 pounds of sugar for a man who raped a black slave, and a white woman who had sexual relations with a black slave was to be fined, imprisoned, and then deported.

Sexual relations between masters and slaves

The labour and food regimes were central to almost every slave's life. In societies where the owner's control over his slave was total, such as the Roman Empire or the pre-1830 American South, the law said little or nothing about how long he could work him and whether his slave had a right to food and clothing. In South India the slave owner had an absolute right to whatever labour his slave was capable of rendering. In Muscovy, on the other hand, a slave owner was jailed for forcing his slaves to labour on Sunday. In Judea in 200 BC, in Sicily in 135–32 BC, and on the Nile in AD 46 regulations prescribed the food rations a slave could expect. The Lithuanian Statute of 1588 and the Russians in 1603 and 1649 decreed that slaves had a right to be fed. The Danish Virgin Islands in 1755 prescribed adequate food rations. The Alabama Slave Code of 1852 mandated that the owner had to provide slaves of working age a sufficiency of healthy food, clothing, attention during illness, and necessities in old age.

A major issue was whether the master had to allow the slave to marry and what rights the owner had over slave offspring. In general, a slave had far fewer rights to his offspring than to his spouse. Babylonian, Hebrew, Tibetan-speaking Nepalese Nyinba, Siamese, and American Southern slave owners thought nothing of breaking up both the conjugal unit and the nuclear family. Unexpectedly the 1755 Danish Virgin Islands Reglement prohibited separating minors from their parents. In Muscovy and China, slave owners could sell or will children apart from their parents, but marriages were inviolable.

The slave's right of marriage and offspring

In North America, India, Rome, Muscovy, most of the Islāmic world, and among the Tuareg a fundamental principle was that the slave could not own property because the master owned not only his slave's body but everything that body might accumulate. This did not mean, however, that slaves could not possess and accumulate property but only that their owners had legal title to whatever the slaves had. In a host of other societies, such as ancient and Roman Egypt, Babylonia, Assyria, Talmudic Palestine, Gortyn, much of medieval Germany, Thailand, Mongol and Ch'ing China, medieval Spain, and the northern Nigerian emirates, slaves had the right of property ownership. Some places, such as Rome, allowed slaves to accumulate, manage, and use property in a peculium that was legally revocable but could be used to purchase their freedom. This provision gave slaves an incentive to work as well as the hope of eventual manumission.

The slave's right of property

Considerable research has been done on the treatment of slaves, and the consensus is that, while the law may have spelled out the desired social standards of master–slave relations, it did not necessarily define the reality for any particular situation. Sadists, even psychopaths, who could not cope with their right of total dominance over another human being, might appear anywhere, as might kindly masters. More determining than the law were the conditions of the society itself. At one extreme, among the Tuareg of North Africa, the slave owners themselves often lived badly, and so, of course, did their slaves. At the other extreme, in the American South material conditions were sufficiently favourable to provide comparative comfort for both masters and slaves. Moreover, slaves born of already enslaved parents usually were treated much better than those purchased or captured from foreign groups. The treatment of slaves in expansive, dynamic societies was likely to be worse than in more stable ones.

Legal relationships between slave owners. There was more uniformity across systems regarding legal relationships between slave owners. All societies had provisions for the recovery of runaways, and most imposed sanctions on owners who stole others' slaves (a capital offense in some systems) or helped them to flee. There also were relatively uniform laws about passing slaves from one generation to another.

There was considerable variability among societies in the law of slave transactions. Whereas Roman-law societies had elaborate norms on contracts, Muscovy had essentially none. Whereas legal systems from Babylonia, Athens, Rome, early Germany, China, and Ethiopia to Islāmic societies and Louisiana allowed guarantees by the sellers that slaves would not flee, were free from disease, or had certain skills, no such laws existed in places such as Muscovy.

Legal relationships between slaves and free strangers. Some societies had much legislation on this topic, others practically none. Where the slave was completely dependent on his owner, few laws existed beyond the normal rules governing any form of property; it was the owner's responsibility to recover damages if a third party killed or assaulted either his cow or his slave. The owner, moreover, was held equally or even more responsible for the slave's actions, ranging from homicide to theft, than was the slave himself, for the society desired that the former control his property and there was no assurance that sanctions, especially money fines, could be enforced against slaves.

Homicide of a slave by a stranger was a revealing test of a society's attitude toward the slave. In Mesopotamia and in Islāmic practice the killer of a slave merely had to compensate the owner for the loss of his property. Elsewhere, however, it was different. Roman law introduced the idea in the Lex Cornelia de Sicariis et Veneficis (the dictator Sulla's enactment on murders and poisoners of 81 BC) that a slave was a person and thus that killing a slave could be a crime. That provision found its way into the Code of Justinian. In North America in the period from 1770 to 1830 the killing of a slave was equated in common law with the murder of a white person. Laws were uniformly harsh when a slave killed a stranger who was a freeman.

Attitudes toward homicide of a slave

Some societies did not allow third parties to assault slaves with impunity. In Muscovy, for example, a slave might have honour and could recover from a third party who injured his honour. Societies elsewhere, however, such as the North American Yurok, Tlingit, and other neighbouring Indians, as well as in the American South, explicitly stated that slaves could have no honour, personal status, or prestige. South Carolina law noted that the slave was not "within the peace of the state, and therefore the peace of the state [was] not broken by an assault and battery on him." Conversely, when a slave assaulted a freeman, the latter often recovered from the slave's owner. Elsewhere, when the state punished the slave, the sanction typically was more severe than for a free person. For example, in Ch'ing China a slave was punished one degree more severely than free citizens for offenses against a freeman.

Most societies, such as those in Athens, Rome, Kievan Rus, Thailand, and Louisiana, did not allow slaves to contract independently with third parties, although some allowed the slave to make a contract on his owner's behalf. The brutal deprivation of rights was expressed in the Alabama case *Creswell's Executor* v. *Walter* (1860); the slave, said the court, had "no legal mind, no will which the law can recognize. . . . Because they are slaves, they are incapable of performing civil acts." On the other hand, in a few societies, as in the ancient Middle East, slaves were

allowed to contract with third parties. Roman slaves were allowed to make contracts in regard to third peculium.

A few societies, such as late Assyria and Muscovy, allowed slaves to testify in court, but most did not. It was a rare society that permitted a slave to serve as a witness against his owner, but some societies, such as ancient Nuzi and Muscovy, allowed slaves to testify against, even to sue, third parties. That was particularly likely to be the case when slaves played a major role in the society, because disputes could not be resolved by the freemen alone without resort to evidence provided by slaves.

Laws of manumission. Laws of manumission varied widely from society to society and within societies across time. They are often viewed as the litmus test of a particular society's views of the slave, that is, of the capacities the slave was likely to exhibit as a free human being. Many Islāmic societies, broadly interpreting the Hebrew prescription, generally prescribed that slave owners had to free their slaves after the passage of a number of years, essentially the length of time they considered it took for an "outsider" to become an "insider." Most other societies allowed masters to free their slaves whenever they wished, although there were exceptions. Some legal systems prescribed manumission when the slave adopted the religion of his owner. It is hardly surprising that manumission was more frequent in systems of household slavery, for intimate relations between master and slave soon converted the outsider into an insider. With notable exceptions, such as Athens, Rome, Muscovy, and some circum-Caribbean societies, many societies required manumission after three generations.

Birth was occasionally a route to manumission. In thriving slave systems such as those of the New World, in harsh systems such as those among the Northwest Coast Indians and the medieval Germanic peoples, or even in milder systems such as those of the Chinese and the Muscovites, a slave's offspring simply added to the slave population. But that was not universally the case; African slave societies, such as the Dahomeans of West Africa, the Ashanti of Ghana, or the Azande living between the Congo and the Nile, prescribed that the offspring of slaves should be free, as part of the process of incorporation into a new lineage. Although Islāmic law did not require manumission upon birth, the Qur'ān recommended it, and slave owners were often inclined to follow the religious tenet. The Aztecs freed all children born in slavery except the offspring of traitors. In Thailand emancipation was considered a pious act, and at their death many owners freed their slaves.

Manumission upon birth

The rate of manumission did not necessarily correspond to the legal ease of manumission. It should be noted, however, that in Rome manumission was relatively easy and was widely practiced, even though there was a 5 percent tax on manumission in the Republic, and the Lex Fufia Caninia of 2 BC forbade manumission by testament of more than a fifth to a half of one's slaves, depending on the number owned. In much of sub-Saharan Africa, manumission was common in most periods, and the freed person typically became a kind of relative in a process of assimilation. In Neo-Babylonia, in Late and Middle Assyria, and in Muscovy manumission was easy but rare; in the American South manumission was comparatively difficult and almost never happened after the prohibition on importing new slaves. The factors of institutional dynamism, expansionism, and profitability, as well as race (see below), may have been the most crucial variants for the South, where manumission was even forbidden in South Carolina in 1820, Mississippi in 1822, Arkansas in 1858, and Maryland and Alabama in 1860; other factors were at work in the ancient Middle East and Muscovy.

Rates of manumission

There was considerable variation among societies as to whether a slave was allowed to accumulate property that he might keep after manumission. One form of such accumulation was the Roman peculium, which legally belonged to the master. One of its heirs was called *coartación*, the self-purchase system, widely used 1,500 years later in Latin America.

After manumission, most societies prescribed a period of legal transition to freedom. In the Roman Empire, China, and elsewhere, this period took three generations and might mean that the grandchild of a slave owner (the "patron") was legally responsible for the grandchild of a slave (his "client"). Thereafter the descendants of the freedman became full members of society, although perhaps still despised. The reason for the legally mandated period of transition to freedom was clear: the slave initially was not a member of the society but an outsider (see below), and it took time to become integrated into the new society. Equally important, the slave was dependent on his owner, and it took time for the freedman and his heirs to become fully self-reliant members of society. If the slave owner and his heirs were not responsible for the freedmen, the fear was, as expressed in the Louisiana Slave Code of 1824, that the latter might otherwise become public wards.

THE SOCIOLOGY OF SLAVERY

The slave as outsider. The slave generally was an outsider. He ordinarily was of a different race, ethnicity, nationality, and religion from his owner. The general rule, as enunciated by the specialist on classical slavery Moses I. Finley, was that "no society could withstand the tension inherent in enslaving its own members." In most cases, the slave was an outsider because he was enslaved against his will in one society and then taken by force to another.

As with nearly all rules, there were exceptions, however. Korea, for reasons that are not understood, was one. India was another exception, because of ritual requirements that the social origins of intimate associates be known; there slaves were ritually distanced from their owners. Muscovite Russia, which had outsider slaves as well, was yet another exception, perhaps because the boundaries between insiders and outsiders were blurred. A number of scholars have pointed out that, although the status of the slaves was uniformly lower than that of comparable free people in every society, the material and sometimes other conditions of slaves were frequently better than those of free people; thus it is not surprising that free people occasionally volunteered to be slaves. What is somewhat more surprising is that so few societies found that form of social welfare to be acceptable; most took measures to prohibit or inhibit it. Solon in 594 BC, for example, forbade enslavement for debt in Athens, and the Lex Poetelia Papiria did the same for Rome, *c.* 326 BC. Muscovy in 1597 prevented self-sale into slavery from becoming hereditary by mandating manumission of such slaves on their owners' deaths.

Regardless of the slave's origin, he was nearly always a marginal person in the society in which he was enslaved. In Africa slaves were despised, and their low status, which was passed on to freedmen, persists to the present time. In most societies most slaves were at the very bottom of society.

Attitudes toward slavery: the matter of race. Slaves in most societies were despised. This is best seen in the homology for slaves. The favourite homology was the woman or wife, then the minor child or an animal. Other terms for slaves were the apprentice, the pauper, the harlot, the felon, the actor, and the complex image of the Southern "Sambo" or Caribbean "Quashee." Throughout history slaves have often been considered to be stupid, uneducable, childlike, lazy, untruthful, untrustworthy, prone to drunkenness, idle, boorish, lascivious, licentious, and cowardly. In China slaves were considered to be "mean" and "base"; in India they were fed table scraps.

The attitudes of the world's great religions toward slavery are of special interest. The Judeo-Christian-Islāmic tradition has been the most tolerant of slavery. Judaic and Islāmic canonical texts refer frequently to slavery and treat it as a natural condition that might befall anyone. But they view it as a condition that should be gotten over quickly. Islāmic practice was based on the assumption that the outsider rapidly became an insider and consequently had to be manumitted after six years. New Testament Christianity, on the other hand, had no prescriptions that slaves be manumitted. Canon law sanctioned slavery. This was attributable at least partially to Christianity's primary focus on spiritual values and salvation after death rather than on temporal conditions and the present life. Under such a regime it mattered little whether someone was a slave or a free person while living on earth.

Christianity's tolerance of slavery

A major issue in the topic of attitudes toward slavery is that of race. Although slaves were usually outsiders and often despised, there nevertheless were different kinds of outsiders and different degrees of contempt. Studies have shown that race made a difference. In Rome, where most owners and slaves were white, manumission was frequent. In Africa, where most owners and slaves were black, lineage incorporation was the primary purpose of slavery, and in most societies slaves were allowed to participate in many aspects of social life. In the American South, however, where the owners were of northern European stock and the slaves of African stock, the degree of social isolation of and contempt for slaves was extraordinary. Southern slaves were forbidden to engage in occupations that might demonstrate their capacities, intermarriage almost never occurred, and manumission was almost unheard of as the reigning publicists proclaimed ever more loudly that blacks lacked any capacity to maintain themselves as free individuals.

Slave occupations. Throughout history the range of occupations held by slaves has been nearly as broad as that held by free persons, but it varied greatly from society to society. The actual range did not depend upon whether the slave lived in a slave-owning or a slave society, although the greatest restrictions appeared in the latter.

To start at the top, the highest position slaves ever attained was that of slave minister, or *ministerialis*. *Ministeriales* existed in the Byzantine Empire, Merovingian France, 11th-century Germany during the Salian dynasty, medieval Muscovy, and throughout the Ottoman Empire. A few slaves even rose to be monarchs, such as the slaves who became sultans and founded dynasties in Islām.

At a level lower than that of slave ministers were other slaves, such as those in the Roman Empire, the Central Asian Samanid domains, Ch'ing China, and elsewhere, who worked in government offices and administered provinces. Some of those slaves were government property, whereas others belonged to private individuals who employed them for government work.

On a level similar to that of slaves working in government were the so-called temple slaves. They were employed by religious institutions in Babylonia, Rome, and elsewhere. Unless they were ultimately destined for sacrifice to the gods, temple slaves usually enjoyed a much easier life than other slaves. They served in occupations ranging from priestess to janitor.

Slaves fought as soldiers and usually were considered of high status. In some societies military slaves belonged to private individuals, in others to the government. In 16th-century Muscovy, for example, cavalrymen purchased slaves who fought alongside them on horseback; in the later 17th century Muscovite slaves were relegated to guarding the baggage train. A special type of slave soldier was the Ottoman janissary. The Islāmic Ottoman Turks confiscated Christian children (called "the tribute children"), took them to Istanbul, and raised them to be professional soldiers, or janissaries. Some janissaries served as members of the palace guard and became involved in the succession struggles of the Ottoman Empire. The Egyptian Mamlūks were also professional soldiers of slave origin who rose to run the entire country. The African Hausa of Zaria and most Sudanic regimes included slaves in all ranks of the soldiery and command. The canoe crews of the West African coast were usually slaves. The British even had detachments of slave soldiers in the Caribbean.

The Ottoman janissary

Societies that explicitly refused to employ slaves in combat, such as Athens in its fleet, Rome in its infantry legions, or the American South in the Civil War, were rare. They took such action because fighting was done by freemen, and it was feared that it would be necessary to free the slaves if they could fight. In fact, all of those slave societies occasionally resorted to using slave soldiers when their military situations became desperate.

In many societies slaves were employed as estate managers or bailiffs. This was especially likely to be the case when it was deemed unfitting for freemen to take or give orders involving other freemen. Where such cultural taboos existed, managers were almost always either real outsiders (imported foreigners) or fictive outsiders (slaves). In Muscovy estate managers were a special category of slave, and they were the first whose registration with the central authorities was required.

Slave estate managers

Still other high-status slaves worked as merchants. Before the invention of the corporation, using slaves was one way to expand the family firm. The practice seems to have begun in Babylonia and was perpetuated in Rome, Spain, the Islāmic world, China, and Africa. Slaves were entrusted with large sums of money and were given charge of long-distance caravans. A few slaves in Muscovy were similarly employed in the Siberian fur trade. Other societies, particularly in the American South, forbade slaves to engage in commerce out of fear that they would sell stolen goods.

In nearly all societies possessing slaves, some slaves were found in what might be termed urban occupations ranging from petty shopkeepers to craftsmen. In the Tredegar Iron Works of Richmond, Va., much of the labour force consisted of slaves. In the American South, ancient Rome, Muscovy, and many other societies, slaves worked as carpenters, tailors, and masons. In Bursa, Tur., some of the finest weaving ever done was by slave craftsmen, who often contracted to fulfill a certain amount of work in exchange for emancipation. The stereotype that slaves were careless and could only be trusted to do the crudest forms of manual labour was disproved countless times in societies that had different expectations and proper incentives.

Only a small portion of slaves throughout history were fortunate enough to be employed in elite or prestige occupations. Most were assigned to strictly physical labour, sometimes the most degrading a society had to offer.

Among the worst forms of slave employment were prostitution and occupations demanding hard physical labour. Mining, often conducted in dangerous conditions causing high death rates, seems to have been the worst. The silver mines at Laurium employed as many as 30,000 slaves, who contributed to the prosperity on which Athenian democracy was based. Slaves were also used in gold mining in Africa and in gold and silver mining in Latin America. Gold and coal mining employed (and killed) millions of state slaves of the Gulag in the Soviet Union between the 1920s and 1956. Slaves have been used on great construction projects such as military fortifications, roads, irrigation projects, and temples from Babylonian to Soviet times. Timber felling for lumber and firewood was another form of hard slave labour, as in the Gulag. Yet another form of brutal slave labour was rowing in the galleys, particularly those that belonged to the Ottoman Empire and sailed the Mediterranean. Tens of thousands of Slavs, victims of Crimean Tatar slave raids, first suffered a hellish existence in the Crimea itself and then ended their days rowing on Ottoman triremes.

Galley slaves

Large numbers of slaves were employed in agriculture. As a general rule, slaves were considered suitable for working some crops but not others. Slaves rarely were employed in growing grains such as rye, oats, wheat, millet, and barley, although at one time or another slaves sowed and especially harvested all of these crops. Most favoured by slave owners were commercial crops such as olives, grapes, sugar, cotton, tobacco, coffee, and certain forms of rice that demanded intense labour to plant, considerable tending throughout the growing season, and significant labour for harvesting. The presence or absence of such crops and their relative profitability were among the major determinants of whether or not a slave-owning society became a slave society. In the Roman Empire employment in olive groves and vineyards occupied many slaves. Sugar cultivation made 9th-century Iraq into a slave society. Rice, coconut, coffee, clove, kola nut, peanut, and sesame cultivation were central occupations in some African societies.

The great discovery in Brazil in the second half of the 16th century was the gang labour system, which was so cost-effective that it made Brazilian sugar cheaper in Europe than the sugar produced in the islands off Africa. A plantation using gang labour could produce, on average, 39 percent more output from comparable inputs than could free farms or farms employing non-gang slave labour. The secret of success was that slaves could be driven, whereas free labour could not; this led to the creation of very

profitable gangs of slaves supervised by white overseers and black drivers. Tobacco and coffee cultivation also used gang labour, but cultivation of these crops was less physically demanding than that of sugar and cotton and led to much lower mortality rates than did sugar and rice.

Domestic service and concubinage

Throughout history domestic service was probably the major slave occupation. Drawing water, hewing wood, cleaning, cooking, waiting on table, taking out the garbage, shopping, child-tending, and similar domestic occupations were the major functions of slaves in all slave-owning societies. In a major productive slave system, the Roman Empire at the time of Augustus and later, the richest 5 percent of Italy's population owned 1,000,000 house slaves (another 2,000,000 were employed elsewhere, out of a total population of about 7,500,000 people). In yet another productive slave system, the American South, large numbers of slaves also worked in their owners' houses. A related function was concubinage, unquestionably one of the major uses of female slaves since the beginning of the institution and particularly prevalent in China. Some societies prescribed that a concubine who bore her owner children was to be freed; others, ranging from the ancient Middle East to the European Middle Ages, specified that the offspring of free–slave unions were to be freed. Rome and the American South were unusual in believing that all concubines and offspring should remain enslaved. Added to this in Africa was the function of lineage expansion, one of the major purposes of slavery in the sub-Saharan region.

Slave marriages

Slave demography. It is sometimes alleged that slavery and marriage were totally incompatible, for recognition of the husband–wife bond would have limited intolerably the slave owner's authority and his right to dispose of his property. Historically, however, such a view is incorrect. Limitations on the right to dispose of property have been frequent throughout history, and slaves were no exception. Thus, slave marriages were recognized in a number of slave-owning societies, including Carthage, Hellenistic Greece, late Byzantium, most of the Roman Catholic medieval world, Ch'ing China, Hindu India, Thailand, the Tlingit and Kwakiutl, and Oregon coast tribes. Ḥanbalī Muslims stated that a slave could insist that his master provide him with a spouse, and Ming Chinese masters were obliged to choose mates for their female slaves when the latter were in their teens and for males around the age of 20. In Russia marriage between a free person and a slave was recognized legally, but according to one of the oldest Russian laws the free person became enslaved by marrying a slave. In Muscovy if a married slave fled, remarried, and was subsequently apprehended, he was to be rejoined to the first spouse.

In the majority of slave societies (the Danish Virgin Islands excepted), on the other hand, slave marriages were not recognized in law and were not something that slave owners had to think about legally when disposing of slaves. For example, the Louisiana Code of 1824 explicitly stated that a slave had no right to be married. Nevertheless, even in these societies, including Rome, the American South, and West Indian Barbuda, slaves formed what they considered marriages and had children. Southern slave owners often recognized such marriages (even across estate boundaries) and their offspring because to have done otherwise would have interfered with production. In Brazil slave marriages were recognized by the Roman Catholic Church and recognized by law in 1869, but in 1875 only one-sixth of the slaves of marriageable age were recorded as married or widowed.

Slave demography was frequently determined by the occupational employment of the slaves. Consequently, sexual imbalance was not at all unusual. In 9th-century France on the Abbey of Saint Germain des Prés' territory there were nearly three male slaves for every female, presumably because of the demand for agricultural labourers. In late medieval Europe, on the other hand, there was a great demand for female slaves as domestics and concubines. The same was true in China, where by the end of the Ch'ing era the institution of slavery had become primarily a female one. In early modern Russia there were two male slaves for every female because of a market demand for cavalrymen, military body servants, and domestics who could perform heavy labour. Concubinage, moreover, was illegal, and those who sold themselves into slavery practiced female infanticide before selling themselves. In many parts of Africa the demand was primarily for women and children for the purpose of incorporation into and expansion of lineages. Adult males were often killed unless they could be exported abroad. Such export conveniently fit into the circum-Caribbean demand for productive slaves to work in sugar, tobacco, and cotton production. Consequently, twice as many males as females and relatively few children under age 10 were shipped to the New World.

One of the notions about slavery has been that slaves rarely reproduced themselves in bondage. Given the skewed demographic profile of many slave societies, it is not surprising that they failed to do so. The slaves of the Athenian Laurium silver mines or the Cuban sugar plantations, for example, lived in largely male societies. In Islāmic slave-owning societies, castration and infibulation curtailed slave reproduction.

The major exception to the rule was North America, where slaves began to procreate in significant numbers in the mid-18th century. This fact helped the slave owners survive the cutting off of imports in 1808. Between the censuses of 1790 and 1860 the slave population of the South expanded enormously—from 657,327 to 3,838,765—one of the fastest rates of population growth ever recorded prior to the advent of modern medicine. Paradoxically, although the Southern slave regime was one of the most dehumanizing ever recorded, it was one of the most favourable on record demographically, because the nutritional and general living environments were highly conducive to explosive population growth. Without significant imports the Southern slave population increased fourfold between the early 1800s and 1860.

Slave population growth

The ages of slave populations also were determined partially by productive requirements. As mentioned above, in Africa children were preferred for incorporation into lineages, whereas in much of the circum-Caribbean world adults were demanded for production. As a consequence, the age pyramids of both societies were skewed; in Africa children predominated, in much of the New World people over age 15. In Muscovy, to take another example, the age structure was skewed toward young adults, for it was primarily young adult males (aged 15–25) who sold themselves into slavery.

Slave protest. Throughout history human beings have objected to being enslaved and have responded in myriad ways ranging from individual shirking, alcoholism, flight, and suicide to arson, murdering owners, and mass rebellion. Perhaps the most common individual response to enslavement was sluggishness, passivity, and indifference. A nearly universal stereotype of the slave was of a lying, lazy, dull brute who had to be kicked or whipped. There probably were three mutually reinforcing factors at work: an unconscious response to overcontrol and absence of freedom, a conscious effort to sabotage the master's desires, and a conditioned response to the expectation of stereotypical behaviour. Some owners tried to overcome such behaviour by a system of incentives or by strict regimentation, such as the gang system, but historically they were in a minority. Less frequent was suicide. A number of slaves are known to have jumped overboard during the Middle Passage because they feared that the transatlantic voyage was taking them to be eaten by witches or barbarians, a fate that seemed worse than drowning.

Flight, either individually or in groups, was one of the most visible forms of protest against enslavement. The rates of flight, which varied greatly from society to society throughout history, usually depended less on individual slave-owner conduct than on the likelihood of success. Immediate conditions, such as the brutality of an overseer or master or a temporary lapse of supervision, often precipitated slave flight, but willingness to undertake such a form of rebellion against the system was usually determined by such factors as the accessibility of refuge or the ability to blend in with the free population (some societies marked slaves to inhibit such blending). Slave flight was infrequent in societies such as the peacetime American South or in West Africa, where a refuge of freedom was

Slave flight

very distant. In East Africa, where flight was curtailed by slave owners united in their desire to prevent it in spite of a high demand for labour, runaways joined neighbouring communities and then raided their former masters. For more than two centuries fugitive slaves in Brazil known as maroons set up independent polities, or *quilombos*, that lasted for years. Maroon communities were found in many other places in Latin America and the Caribbean as well. In Muscovy, where most of the slaves were natives or of similar origin (Poles and Swedes), where there was an open frontier, and where masters had no compunction about taking in other owners' slaves, the rate of flight was very high; and as many as a quarter to a third of the slaves ran away. In China flight by male slaves was also common. During the American Revolution, when the slave owners were occupied with fighting the British, fugitive slaves numbered in the tens of thousands.

Direct, personal attacks on slave owners often were determined by the nature of the slave regime. Where owners believed they enjoyed automatic sexual access to female slaves, both the women and their "husbands" were prone to respond by assaulting the owners or their agents. In Hausaland, killings by concubines instilled great fear in slave owners. Where slaves were driven, assault on the drivers was not an uncommon response. As a result, overseers in the Mississippi Valley feared for their lives and constantly carried arms.

The most dramatic form of slave protest was outright rebellion. Slave uprisings varied enormously in frequency, size, intensity, and duration. Perhaps the calmest of all known slave societies were those of West Africa, where the predominance of women and children caused rebellions to be very few. Slave rebellions in North America were also noticeably few and involved only a handful of participants: the New York revolt of 1712, the Stono rebellion of South Carolina (1739), the Gabriel plot in Richmond, Va. (1800), the Denmark Vesey conspiracy in Charleston, S.C. (1822), and Nat Turner's uprising in Jerusalem, Va. (1831), are the best known. Southern slave uprisings were so few and so small because of the absolute certainty that they would be brutally repressed. The Turner rebellion is usually given as the reason for the marked increase in the severity of the slave regime after 1831.

Slave uprisings

Other slave revolts were on a much grander scale than those of West Africa and North America. One of the most famous slave uprisings was the Gladiatorial War led by Spartacus against Rome in 73–71 BC. The Spartacus rebellion was brutally repressed (the roads leading into Rome were lined with gibbets from which rebel corpses hung). Slaves led the Khlopko and Bolotnikov uprisings in Muscovy in 1603 and 1606, respectively, a time of dynastic crisis. Another great slave rebellion was that of the Zanj (black slaves imported from Zanzibar) in Iraq and Khuzistan in the years 869–883. It was joined by fair-skinned slaves as well and was on a larger scale than the Spartacus revolt. Slave rebellion in China at the end of the 17th and the beginning of the 18th century was so extensive that owners eventually eschewed male slaves and converted the institution into a female-dominated one.

Slave rebellions occurred in every slave society in the Americas from the 16th century onward. Prominent slave revolts occurred in Jamaica in 1760, 1798, and 1831–32, in Barbados in 1816, and in British Guiana in 1823. Perhaps the most famous Caribbean rebellion, in Saint-Domingue, began in 1791 and was subsequently led to victory by the freedman Toussaint-Louverture; it produced the emancipation of its slaves while the French were preoccupied with their own revolution and ultimately led to the independent state of Haiti.

SLAVE CULTURE

The institution of slavery usually tried to deny its victims their native cultural identity. Torn out of their own cultural milieus, they were expected to abandon their heritage and to adopt at least part of their enslavers' culture. Nonetheless, studies have shown that there were aspects of slave culture that differed from the master culture. Some of these have been interpreted as a form of resistance to oppression, while other aspects were clearly survivals of a native culture in the new society. Most of what is known about this topic comes from the circum-Caribbean world, but analogous developments may have occurred wherever alien slaves were concentrated in numbers sufficient to prevent their complete absorption by the host slave-owning or slave society. Thus slave culture was probably very different on large plantations from what it was on small farms or in urban households, where slave culture (and especially Creole slave culture) could hardly have avoided being very similar to the master culture. Slave cultures grew up within the perimeters of the masters' monopoly of power but separate from the masters' institutions.

Religion, which performed the multiple function of explanation, prediction, control, and communion, seems to have been a particularly fruitful area for the creation of slave culture. Africans perceived all misfortunes, including enslavement, as the result of sorcery, and their religious practices and beliefs, which were often millennial, were formulated as a way of coping with it. Myalism was the first religious movement to appeal to all ethnic groups in Jamaica, Voodoo in Haiti was the product of African culture slightly refashioned on that island, and syncretic Afro-Christian religions and rituals appeared nearly everywhere throughout the New World. Slave religions usually had a supreme being and a host of lesser spirits brought from Africa, borrowed from the Amerindians, and created in response to local conditions. There were no firm boundaries between the secular and the sacred, which infused all things and activities. At least initially African slaves universally believed that posthumously they would return to their lands and rejoin their friends.

African elements in New World slave cultures

Black slaves preserved some of their culture in the New World. African medicine was practiced in America by slaves. The poisoning of masters and other hated individuals was a particularly African method of coping with evil. Throughout the circum-Caribbean world slaves and free blacks had electoral procedures, adapted from West African customs, to choose governors, sheriffs, and judges to maintain order among themselves. Objects of material culture, such as rugs, mats, baskets, thatched roofs, and walking canes, were modeled on African examples. Nevertheless, relatively few African social practices or plastic arts survived in the New World. On the other hand, Afro-American music and dance are known to have many African roots, and they differed dramatically from the practices of the European master culture; the use of drum and banjo were especially significant. Songs and spirituals borrowed their strong call-and-response patterns from the West African style. Furthermore, slaves created tales to amuse themselves, and the African element is most evident in animal tales; the tar-baby story is among the best known of the genre. Afro-American stories and songs often featured the devil, who was a demon and a trickster, terrifying, a friend in need, and a source of mirth.

Slave culture also developed beliefs and customs that were at odds with those of the master culture. One such belief was that what the masters called theft was something else; thus stealing from the master was not theft at all but merely a process of channeling his property from one use to another, as in taking his corn and feeding it to his pigs. Polygamous domestic arrangements were a further aspect of slave culture brought from Africa. Yet another aspect of slave culture, especially prevalent in the Caribbean, involved the market. Slaves there were often required to provide their own food, which they raised on provision grounds. If they had any surplus, they were permitted by their owners to sell it in the market. As a result slaves developed an autonomy and an individualism that contrasted starkly with the rigid control of the work gang system and the putative stifling control of slave law.

BIBLIOGRAPHY

General works: Four specialized encyclopedias collectively provide comprehensive coverage and ready access to every aspect of the subject: JUNIUS P. RODRIGUEZ (ed.), *The Historical Encyclopaedia of World Slavery*, 2 vol. (1997); SEYMOUR DRESCHER and STANLEY L. ENGERMAN (eds.), *A Historical Guide to World Slavery* (1998); PAUL FINKELMAN and JOSEPH C. MILLER (eds.), *Macmillan Encyclopedia of World Slavery*, 2 vol. (1998); and JUNIUS P. RODRIGUEZ, *Chronology of World Slavery* (1999). A

treatment of the economics of forced labour, including slavery, is M.L. BUSH (ed.), *Serfdom and Slavery: Studies in Legal Bondage* (1996). Another wide-ranging collection of essays in this vein is STANLEY L. ENGERMAN (ed.), *Terms of Labor: Slavery, Serfdom, and Free Labor* (1999). For the philosophy and ethics of slavery within the European tradition, DAVID BRION DAVIS, *The Problem of Slavery in Western Culture* (1966, reissued 1988), *The Problem of Slavery in the Age of Revolution, 1770–1823* (1975), and *Slavery and Human Progress* (1984), may be consulted.

Slavery in early history: The classic monographs on slavery in ancient empires are ISAAC MENDELSOHN, *Slavery in the Ancient Near East: A Comparative Study of Slavery in Babylonia, Assyria, Syria, and Palestine, from the Middle of the Third Millennium to the End of the First Millennium* (1949, reprinted 1978); and MUHAMMAD A. DANDAMAEV (M.A. Dandamaev), *Slavery in Babylonia: From Nabopolassar to Alexander the Great (626–331 B.C.)*, rev. ed., edited by MARVIN A. POWELL and DAVID B. WEISBERG (1984; originally published in Russian, 1974).

Central works on slavery in classical antiquity are M.I. FINLEY, *Slavery in Classical Antiquity: Views and Controversies* (1960), containing his key shorter writings, *The Ancient Economy*, updated ed. (1999), and his general historiographical essay covering the background to the vast literature in this field in European languages, the classic *Ancient Slavery and Modern Ideology*, expanded ed., edited by BRENT D. SHAW (1998). Later survey interpretations are T.E.J. WIEDEMANN, *Slavery: with Addenda (1992) and Further Addenda (1997)* (1997); and PETER GARNSEY, *Ideas of Slavery from Aristotle to Augustine* (1996). Works that focus on Greece include YVON GARLAN, *Slavery in Ancient Greece* (1988; originally published in French, 1982); and PETER HUNT, *Slaves, Warfare, and Ideology in the Greek Historians* (1998). Works specifically on Rome are KEITH BRADLEY (K.R. Bradley), *Slavery and Society at Rome* (1994); and, on Roman law, ALAN D. WATSON, *Roman Slave Law* (1987).

Slavery in Islāmic lands: Works that focus on the Islāmic world are MURRAY GORDON, *Slavery in the Arab World* (1989, reissued 1992; originally published in French, 1987); and SHAUN E. MARMON (ed.), *Slavery in the Islamic Middle East* (1999). Military slaves are the subject of DAVID AYALON, *Islam and the Abode of War: Military Slaves and Islamic Adversaries* (1994), which reprints classic studies; formative also has been DANIEL PIPES, *Slave Soldiers and Islam: The Genesis of a Military System* (1981); contemporary views are collected in MIURA TORU and JOHN EDWARD PHILIPS (eds.), *Slave Elites in the Middle East and Africa: A Comparative Study* (2000). The Ottoman Empire is discussed in EHUD R. TOLEDANO, *Slavery and Abolition in the Ottoman Middle East* (1998); and Y. HAKAN ERDEM, *Slavery in the Ottoman Empire and Its Demise, 1800–1909* (1996). The supporting trades from Africa are discussed in EHUD R. TOLEDANO, *The Ottoman Slave Trade and Its Suppression, 1840–1890* (1982); WILLIAM GERVASE CLARENCE-SMITH (ed.), *The Economics of the Indian Ocean Slave Trade in the Nineteenth Century* (1989); and ELIZABETH SAVAGE (ed.), *The Human Commodity: Perspectives on the Trans-Saharan Slave Trade* (1992).

Slavery in Europe: A good general synthesis is WILLIAM D. PHILLIPS, JR., *Slavery from Roman Times to the Early Transatlantic Trade* (1985). The foundational work was MARC BLOCH, *Slavery and Serfdom in the Middle Ages: Selected Essays* (1975; originally published in French, 1963). The subject is amply documented in CHARLES VERLINDEN, *L'esclavage dans l'Europe médiévale* (1955–). Major interpretations include PIERRE DOCKÈS, *Medieval Slavery and Liberation* (1982; originally published in French, 1979); JACQUES HEERS, *Esclaves et domestiques au Moyen Âge dans le monde meditérranéen* (1981, reissued 1996); and PIERRE BONNASSIE, *From Slavery to Feudalism in South-Western Europe*, trans. from French (1991).

Slavery in Asia: Collections of essays include JAMES L. WATSON (ed.), *Asian and African Systems of Slavery* (1980); ANTHONY REID and JENNIFER BREWSTER (eds.), *Slavery, Bondage, and Dependency in Southeast Asia* (1983); and MARTIN A. KLEIN (ed.), *Breaking the Chains: Slavery, Bondage, and Emancipation in Modern Africa and Asia* (1993). India is the focus of UTSA PATNAIK and MANJARI DINGWANEY (eds.), *Chains of Servitude: Bondage and Slavery in India* (1985); GYAN PRAKASH, *Bonded Histories: Genealogies of Labor Servitude in Colonial India* (1990); and INDRANI CHATTERJEE, *Gender, Slavery and Law in Colonial India* (1999).

Slavery in sub-Saharan Africa: General interpretations include PAUL E. LOVEJOY, *Transformations in Slavery: A History of Slavery in Africa*, 2nd ed. (2000); and PATRICK MANNING, *Slavery and African Life: Occidental, Oriental and African Slave Trades* (1990). The internal aspects of slavery within Africa are discussed in SUZANNE MIERS and IGOR KOPYTOFF (eds.), *Slavery in Africa: Historical and Anthropological Perspectives* (1977); CLAIRE C. ROBERTSON and MARTIN A. KLEIN (eds.), *Women and Slavery in Africa* (1983); SUZANNE MIERS and RICHARD ROBERTS (eds.), *The End of Slavery in Africa* (1988); and SUZANNE MIERS and MARTIN KLEIN (eds.), *Slavery and Colonial Rule in Africa* (1998). One of many sources that treat the role of the Dutch is ELIZABETH A. ELDREDGE and FRED MORTON (eds.), *Slavery in South Africa: Captive Labor on the Dutch Frontier* (1994).

Atlantic slave trade: All known voyages carrying slaves to the New World (27,233) are accessible in DAVID ELTIS *et al.* (eds.), *The Trans-Atlantic Slave Trade: A Database on CD-ROM* (1999), with an extensive introduction. The foundational work for demographic study of the trade is PHILIP D. CURTIN, *The Atlantic Slave Trade: A Census* (1969); a comprehensive update in this vein is HERBERT S. KLEIN, *The Atlantic Slave Trade* (1999).

Slavery in the New World: One of the rare studies to focus on Native American slavery is LELAND DONALD, *Aboriginal Slavery on the Northwest Coast of North America* (1997). General interpretations of the Atlantic slave trade and African slavery in the Americas begin with ERIC WILLIAMS, *Capitalism & Slavery* (1944, reissued 1994); a late 20th-century offering in this vein is ROBIN BLACKBURN, *The Making of New World Slavery: From the Baroque to the Modern* (1997). A hemispheric interpretation of slave revolts is EUGENE D. GENOVESE, *From Rebellion to Revolution: Afro-American Slave Revolts in the Making of the Modern World* (1979, reissued 1992); and a comprehensive assessment of slaves' contribution to emancipation is ROBIN BLACKBURN, *The Overthrow of Colonial Slavery, 1776–1848* (1988). An older source is RICHARD PRICE (ed.), *Maroon Societies: Rebel Slave Communities in the Americas*, 3rd ed. (1996). HERBERT S. KLEIN, *African Slavery in Latin America and the Caribbean* (1986, reissued 1988), is an overview of slavery south of the U.S. border. For the Spanish-speaking mainland cultures, one finds only scattered monographic work in English.

The literature on slavery in North America is vast. An up-to-date general survey is PETER KOLCHIN, *American Slavery, 1619–1877* (1993, reissued 1995). Two major interpretations of slavery in the colonial era are IRA BERLIN, *Many Thousands Gone: The First Two Centuries of Slavery in North America* (1998); and PHILIP D. MORGAN, *Slave Counterpoint: Black Culture in the Eighteenth-Century Chesapeake and Lowcountry* (1998); an older treatment is WINTHROP D. JORDAN, *White over Black: American Attitudes Toward the Negro, 1550–1812* (1968, reissued 1977).

Classics in the interpretation of slavery in the American South, when cotton was king, start with ULRICH B. PHILLIPS, *American Negro Slavery: A Survey of the Supply, Employment, and Control of Negro Labor as Determined by the Plantation Régime* (1918, reissued 1966); which should be read with JOHN DAVID SMITH and JOHN C. INSCOE (eds.), *Ulrich Bonnell Phillips: A Southern Historian and His Critics* (1990). Other notable works are KENNETH M. STAMPP, *The Peculiar Institution: Slavery in the Ante-Bellum South* (1956, reissued 1995); JOHN W. BLASSINGAME, *The Slave Community: Plantation Life in the Antebellum South*, rev. ed. (1979); ROBERT WILLIAM FOGEL and STANLEY L. ENGERMAN, *Time on the Cross: The Economics of American Negro Slavery* (1974, reissued 1989); EUGENE D. GENOVESE, *Roll, Jordan, Roll: The World the Slaves Made* (1974); CLAUDIA DALE GOLDIN, *Urban Slavery in the American South, 1820–1860* (1976); STERLING STUCKEY, *Slave Culture: Nationalist Theory and the Foundations of Black America* (1987); JAMES OAKES, *The Ruling Race: A History of American Slaveholders* (1982); and ROBERT WILLIAM FOGEL, *Without Consent or Contract: The Rise and Fall of American Slavery* (1989, reprinted 1994).

Many studies, most in Portuguese, cover Brazilian slavery. Prominent works in English include GILBERTO FREYRE, *The Masters and the Slaves: A Study in the Development of Brazilian Civilization* (1946; originally published in Portuguese, 1933); ROBERT EDGAR CONRAD, *Children of God's Fire: A Documentary History of Black Slavery in Brazil* (1983, reissued 1994); STUART B. SCHWARTZ, *Sugar Plantations in the Formation of Brazilian Society: Bahia, 1550–1835* (1985), and *Slaves, Peasants, and Rebels: Reconsidering Brazilian Slavery* (1992, reissued 1996); KÁTIA M. DE QUEIRÓS MATTOSO, *To Be a Slave in Brazil, 1550–1888* (1986; originally published in French, 1979); JOÃO JOSÉ REIS, *Slave Rebellion in Brazil: The Muslim Uprising of 1835 in Bahia*, rev. and expanded ed. (1993; originally published in Portuguese); B.J. BARICKMAN, *A Bahian Counterpoint: Sugar, Tobacco, Cassava, and Slavery in the Recôncavo, 1780–1860* (1998); and KATHLEEN J. HIGGINS, *"Licentious Liberty" in a Brazilian Gold-Mining Region: Slavery, Gender, and Social Control in Eighteenth-Century Sabarâ, Minas Gerais* (1999).

The circum-Caribbean region is discussed in such volumes as MICHAEL CRATON, *Empire, Enslavement, and Freedom in the Caribbean* (1997); ROBERT LOUIS STEIN, *The French Sugar Business in the Eighteenth Century* (1988); FRANKLIN W. KNIGHT, *Slave Society in Cuba During the Nineteenth Century* (1970, reissued 1986); FRANCISCO A. SCARANO, *Sugar and Slavery in Puerto Rico: The Plantation Economy of Ponce, 1800–1850* (1984); and JOHANNES POSTMA, *The Dutch in the Atlantic Slave Trade, 1600–1815* (1990).

(R.He./J.C.M/Ed.)

Sleep and Dreams

Sleep is a normal, easily reversible, recurrent, and spontaneous state of decreased and less efficient responsiveness to external stimulation. The state contrasts with that of wakefulness, in which there is an enhanced potential for sensitivity and an efficient responsiveness to external stimuli. The sleep–wakefulness alternation is the most striking manifestation in higher vertebrates of the more general phenomenon of periodicity in the activity or responsivity of living tissue (see BEHAVIOUR, ANIMAL). There is no single, perfectly reliable criterion of sleep. Sleep is defined by the convergence of observations satisfying several different motor, sensory, and physiological criteria. Occasionally, one or more of these criteria may be absent during sleep or present during wakefulness, but even in such cases there usually is little difficulty in achieving agreement among observers in the discrimination between the two behavioral states.

Dreaming, a common and distinctive phenomenon of sleep, has since the dawn of human history given rise to myriad beliefs, fears, and conjectures, both imaginative and experimental, regarding its mysterious nature. While any effort toward classification must be subject to inadequacies, beliefs about dreams fall into various classifications depending upon whether dreams are held to be reflections of reality, sources of divination, curative experiences, or evidence of unconscious activity.

This article treats first the psychological and physiological characteristics of the sleeping state. Next, dreaming is examined from both historical-cultural and scientific points of view.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 422 and 433, and the *Index.*

The article is divided into the following sections:

Sleep 301
- The nature of sleep 301
- Developmental patterns of sleep and wakefulness 302
- Psychophysiological variations in sleep 302
  - Non-rapid eye movement sleep
  - Rapid eye movement sleep
  - Sequences of NREM and REM sleep
- Effects of sleep deprivation 304
  - General sleep deprivation
  - Selective sleep deprivation
- Pathological aspects 305
  - Primary disturbances
  - Minor episodes
  - Disorders accentuated during sleep
  - Disorders of sleep schedule
- Drugs and sleep 306
- Theories of sleep 307
  - Mechanistic theories
  - Neural theories
  - Functional theories

Dreams and dreaming 308
- Diverse views on the nature of dreams 308
  - Dreams as reflecting reality
  - Dreams as a source of divination
  - Dreams as curative
  - Dreams as extensions of the waking state
  - Psychoanalytic interpretations
- Efforts to study dreaming 309
  - Dream reports
  - Physiological dream research
- Dreamlike activities 310

Bibliography 311

Sleep

THE NATURE OF SLEEP

Motor and sensory criteria used in defining sleep

Sleep usually requires the presence of flaccid or relaxed skeletal muscles and the absence of the overt, goal-directed behaviour of which the waking organism is capable. Part of the recurring fascination with sleep talking and sleepwalking stems from their apparent violation of this latter criterion. Were these phenomena continuous, rather than intermittent, during a behavioral state, it is indeed questionable whether the designation "sleep" would continue to be appropriate. The characteristic posture associated with sleep in man and in many but not all animals is that of horizontal repose. The relaxation of the skeletal muscles in this posture and its implication of a more passive role toward the environment are symptomatic of sleep.

Indicative of the decreased sensitivity of the human sleeper to his external environment are the typical closed eyelids (or the functional blindness associated with sleep while the eyes are open) and the presleep activities that include seeking surroundings characterized by reduced or monotonous levels of sensory stimulation. Three additional criteria—reversibility, recurrence, and spontaneity—distinguish the insensitivity of sleep from that of other states. Compared to that of hibernation or coma, the insensitivity of sleep is more easily reversible. Although the occurrence of sleep is not perfectly regular under all conditions, it is at least partially predictable from a knowledge of the duration of prior sleep periods and of the intervals between periods of sleep; and, although the onset of sleep may be facilitated by a variety of environmental or chemical means, sleep states are not thought of as being absolutely dependent upon such manipulations.

In experimental studies, both with subhuman vertebrates and with humans, sleep also has been defined in terms of physiological variables generally associated with recurring periods of inactivity identified behaviorally as sleep. For example, the typical presence of certain electroencephalogram (EEG) patterns (brain patterns of electrical activity as recorded in tracings) with behavioral sleep has led to the designation of such patterns as "signs" of sleep. Conversely, in the absence of such signs (as, for example, in a hypnotic trance) it is felt that true sleep is absent. Such signs as are now employed, however, are not invariably discriminating of the behavioral states of sleep and wakefulness. Advances in the technology of animal experimentation have made it possible to extend the physiological approach from externally measurable manifestations of sleep such as the EEG to the underlying neural (nerve) mechanisms presumably responsible for such manifestations. As a result, it may finally become possible to identify structures or functions that are invariably related to behavioral sleep and to trace the evolution of sleep through comparative anatomic and physiological studies of structures found to be critical in the maintenance of sleep behaviour in the higher vertebrates.

In addition to the behavioral and physiological criteria already mentioned, subjective experience (in the case of the self) and verbal reports of such experience (in the case of others) are used at the human level to define sleep. Upon being alerted, one may feel or say, "I was asleep just then," and such judgments ordinarily are accepted as evidence for identifying a pre-arousal state as sleep, but such subjective evidence can be at variance with behaviouristic classifications of sleep.

Problems in defining sleep

More generally, problems in defining sleep arise when evidence for one or more of the several criteria of sleep is lacking or when the evidence generated by available criteria is inconsistent. Do subhuman species sleep? Other

mammalian species whose EEG and other physiological correlates are akin to those observed in human sleep demonstrate recurring, spontaneous, and reversible periods of inactivity and decreased critical reactivity. There is general acceptance of the designation of such states as sleep. As one descends the evolutionary scale below the birds and reptiles, however, and such criteria are successively less well satisfied, the unequivocal identification of sleep becomes more difficult. Bullfrogs (*Rana catesbeiana*), for example, seem not to fulfill sensory threshold criteria of sleep during resting states. Tree frogs (genus *Hyla*), on the other hand, show diminished sensitivity as they move from a state of behavioral activity to one of rest. Yet the EEGs of the alert rest of the bullfrog and the sleeplike rest of the tree frog are the same. There are parallel problems in defining sleep at different stages in the development of a single individual. At full-term birth in the human being, for instance, a convergence of nonsubjective criteria clearly seems to justify the identification of periods of sleep, but it is more difficult to justify the attribution of sleep to the human fetus.

Problems in defining sleep may arise from the effects of artificial manipulation. For example, the EEG patterns commonly used as signs of sleep can be induced in an otherwise waking organism by the administration of certain drugs. Sometimes, also, there is conflicting evidence: a person who is "awakened" from a spontaneously assumed state of immobility with all the EEG criteria of sleep may claim that he had been awake prior to this event. In such troublesome cases and more generally, it is becoming common to qualify attributions of sleep with the criteria upon which such attributions rest—*e.g.*, "behavioral sleep," "physiological sleep," or "self-described sleep." Such terminology accurately reflects both the multiplicity of criteria available for the identification of sleep and the possibility that these criteria may not always agree with one another.

DEVELOPMENTAL PATTERNS OF SLEEP AND WAKEFULNESS

Length of time spent in sleep

How much sleep does a person need? While the physiological bases of the need for sleep remain conjectural, rendering definitive answers to this question impossible, much evidence has been gathered on how much sleep people do in fact obtain. Perhaps the most important conclusion to be drawn from this evidence is that there is great variability among individuals in total sleep time. For adults, anything between six and nine hours of sleep as a nightly average is not unusual, and 7½ hours probably best expresses the norm. Such norms, of course, inevitably vary with the criteria of sleep employed. The most precise and reliable figures on sleep time, including those cited here, come from studies in sleep laboratories, where EEG criteria are employed.

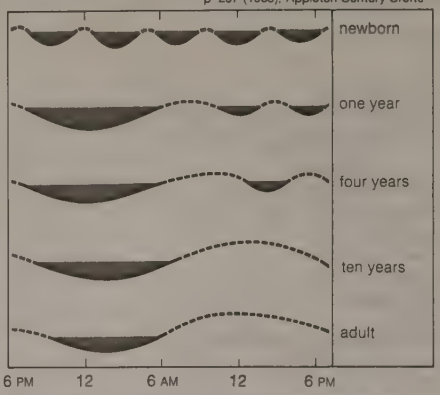
Age consistently has been associated with the varying amount, quality, and patterning of electrophysiologically defined sleep. The newborn infant may spend an average of about 16 hours of each 24-hour period in sleep, although there is wide variability among individual babies. During the first year of life, total sleep time drops sharply; by two years of age, it may range from nine to 12 hours. Decreases to approximately six hours have been observed among the elderly.

As will be elaborated below, EEG sleep studies have indicated that sleep can be considered to consist of several different stages. Developmental changes in the relative proportion of sleep time spent in these sleep stages are as striking as age-related changes in total sleep time. For example, the newborn infant may spend 50 percent of total sleep time in a stage of EEG sleep that is accompanied by intermittent bursts of rapid eye movements (REMs) indicative of a type of sleep that in some respects bears more resemblance to wakefulness than to other forms of sleep (see below *Rapid eye movement sleep*), while the comparable figure for adults is approximately 25 percent, and for the aged is less than 20 percent. There is also a decline with age of EEG stage 4 (deep slumber).

Aspects of sleep patterning

Sleep patterning consists of (1) the temporal spacing of sleep and wakefulness within a 24-hour period and (2) the ordering of different sleep stages within a given sleep period. In both senses, there are major developmental changes in the patterning of sleep. In alternations between sleep and wakefulness, there is a developmental shift from polyphasic sleep to monophasic sleep (*i.e.*, from intermittent to uninterrupted sleep). At birth, there may be five or six periods of sleep per day alternating with a like number of waking periods. With the dropping of nocturnal feedings in infancy and of morning and afternoon naps in childhood, there is an increasing tendency to the concentration of sleep in one long nocturnal period (see the Figure). The trend to monophasic sleep probably reflects some blend of the effects of maturing and of pressures from a culture geared to daytime activity and nocturnal rest. Among the elderly there may be a partial return to the polyphasic sleep pattern of infancy and early childhood, namely, more frequent daytime napping and less extensive periods of nocturnal sleep because of the loss of zeitgebers, time markers that provide cues. These include the need to arise at a set time for work or to get children off to school. Significant developmental effects also have been observed in spacing of stages within sleep. In the adult, REM sleep rarely occurs at sleep onset, while, in newborn infants, sleep-onset REM sleep is typical.

From E.J. Murray, *Sleep, Dreams and Arousal*, p. 297 (1965); Appleton-Century-Crofts



Alternations of sleep and wakefulness at specified ages. The shading indicates periods of sleep.

It would be difficult to overestimate the significance of the various age-related changes in sleep behaviour for a general theory of sleep. In the search for the functional significance of sleep or of particular stages of sleep, the shifts in sleep variables can be linked with variations in waking developmental needs, in the total capacities of the individual, and in environmental demands. It has been suggested, for instance, that the high frequency and priority in the night of REM sleep in the newborn infant may reflect a need for stimulation from within to permit orderly maturation of the central nervous system (CNS). Another interpretation of age-related changes in REM sleep stresses its possible role in processing new information, the rate of acquisition for which is assumed to be relatively high in childhood but reduced in old age. As these views illustrate, developmental changes in the electrophysiology of sleep are germane not only to sleep but also to the role of CNS development in behavioral adaptation.

PSYCHOPHYSIOLOGICAL VARIATIONS IN SLEEP

That there are different kinds of sleep has long been recognized. In everyday discourse there is talk of "good" sleep or "poor" sleep, of "light" sleep and "deep" sleep; yet, only in the second half of the 20th century have scientists paid much attention to qualitative variations within sleep. Sleep was formerly conceptualized by scientists as a unitary state of passive recuperation. Revolutionary changes have occurred in scientific thinking about sleep, the most important of which has been increased sensitivity to its heterogeneity.

This revolution may be traced back to the discovery of sleep characterized by rapid eye movement (REM sleep), first reported by the physiologists Eugene Aserinsky and Nathaniel Kleitman in 1953. REM sleep proved to have characteristics quite at variance with the prevailing model

of sleep as recuperative deactivation of the central nervous system. Various central and autonomic nervous system measurements seemed to show that the REM stage of sleep is more nearly like activated wakefulness than it is like other sleep. It now has become conventional to consider REM ("paradoxical") and non-REM (NREM or "orthodox") sleep as qualitatively different. Thus, the earlier assumption that sleep is a unitary and passive state has yielded to the viewpoint that there are two different kinds of sleep, a relatively deactivated NREM phase and an activated REM phase.

Stages of NREM sleep

Non-rapid eye movement sleep. NREM sleep itself is conventionally subdivided into several different stages on the basis of EEG criteria. In the adult, stage 1 is observed at sleep onset or after momentary arousals during the night and is defined as a low-voltage mixed-frequency EEG tracing with a considerable representation of theta-wave (four to seven hertz, or cycles per second) activity. Stage 2 is a relatively low-voltage EEG tracing characterized by intermittent, short sequences of waves of 12–14 hertz ("sleep spindles") and by formations called K-complexes—biphasic wave forms that can be induced by external stimulation, as by a sound, but that also occur spontaneously during sleep. Stages 3 and 4 consist of relatively high-voltage (more than 50-microvolt) EEG tracings with a predominance of delta-wave (one to two hertz) activity; the distinction between the two stages is based on an arbitrary criterion of amount of delta-wave activity, with greater amounts classified as stage 4. Unlike the basic distinction between NREM and REM, differences among NREM sleep stages generally are regarded as quantitative rather than qualitative.

The EEG patterns of NREM sleep, particularly of stages 3 and 4 (tracings of slower frequency and higher amplitude), are those associated in other circumstances with decreased vigilance. Furthermore, after the transition from wakefulness to NREM sleep, most functions of the autonomic nervous system decrease their rate of activity and their moment-to-moment variability. Thus, NREM sleep is the kind of seemingly restful state that appears capable of supporting the recuperative functions assigned to sleep. There are, in fact, several lines of evidence suggesting such functions for NREM stage 4: (1) increases in such sleep, in both man and laboratory animals, have been observed after physical exercise; (2) the concentration of such sleep in the early portion of the sleep period (*i.e.,* immediately after wakeful states of activity) in human beings; and (3) the relatively high priority that such sleep has, among human beings, in "recovery" sleep following abnormally extended periods of wakefulness.

Rapid eye movement sleep. REM sleep is a state of diffuse bodily activation. Its EEG patterns (tracings of faster frequency and lower amplitude than in NREM stages 2–4) are at least superficially similar to those of wakefulness. Most autonomic variables exhibit relatively high rates of activity and variability during REM sleep; for example, there are higher heart and respiration rates and more short-term variability in these rates than in NREM sleep, increased blood pressure, and, in males, full or partial penile erection. In addition, REM sleep is accompanied by a relatively low rate of gross body motility, but with some periodic twitching of the muscles of the face and extremities, relatively high levels of oxygen consumption by the brain, increased cerebral blood flow, and higher brain temperature. An even more impressive demonstration of the activation of REM sleep is to be found in the firing rates of individual cerebral neurons, or nerve cells, in experimental animals: during REM sleep such rates exceed those of NREM sleep and often equal or surpass those of wakefulness. Another distinguishing feature of REM sleep, of course, is the intermittent appearance of bursts of the rapid eye movements, whence the term is derived.

The three events defining REM sleep

For both humans and animals, REM sleep now is defined by the concurrence of three events: low-voltage, mixed-frequency EEG; intermittent REMs; and suppressed tonus of the muscles of the facial region (*i.e.,* suppression of the continuous slight tension otherwise normally present). This decrease in muscle tonus and a similarly observed suppression of spinal reflexes are indicative of heightened motor inhibition during REM sleep. Animal studies have identified the locus ceruleus, in the pons, as the probable source of this inhibition. (The pons is in the brain stem directly above the medulla oblongata; the locus ceruleus borders on the brain cavity known as the fourth ventricle.) When this structure is surgically destroyed in experimental animals, they periodically engage in active, apparently goal-directed behaviour during REM sleep, although they still show the unresponsivity to external stimulation characteristic of the stage. It has been suggested that such behaviour may be the acting out of the hallucinations of a dream.

An important theoretical distinction is that between REM sleep phenomena that are continuous and those that are intermittent. Tonic (continuous) characteristics of REM sleep include the low-voltage EEG and the suppressed muscle tonus; intermittent events in REM sleep include the REMs themselves and, as observed in the cat, spikelike electrical activity in those parts of the brain concerned with vision and in other parts of the cerebral cortex. The various intermittent events of REM sleep tend to occur together, and it seems to be these moments of intermittent activation that are responsible for much of the difference between REM sleep and NREM sleep. The spiking mentioned is observed occasionally in NREM sleep, an occurrence that has been interpreted by some theorists as suggesting that REM sleep is not qualitatively unique in its capacity to support intermittent activation and that between NREM and REM sleep the differences may be less striking than the differences in eye movement and in EEG have indicated.

Sequences of NREM and REM sleep. The usual temporal progression of the two kinds of sleep in the adult human is for a period of approximately 70–90 minutes of NREM sleep (the stages being ordered 1–2–3–4–3–2) to precede the first period of REM sleep, which may last from approximately five to 15 minutes. NREM–REM cycles of roughly equivalent total duration then recur through the night, with the REM portion lengthening somewhat, and the NREM portion shrinking correspondingly, as sleep continues. Approximately 25 percent of total accumulated sleep is spent in REM sleep and 75 percent in NREM sleep. Most of the latter is EEG stage 2. The high proportion of stage 2 NREM sleep is attributable to the loss of stages 3 and 4 in the NREM portion of the NREM–REM cycles after the first two or three.

Light and deep sleep. Which of the various NREM stages is light sleep and which is deep sleep? The criteria used to establish sleep depth are the same as those used to distinguish sleep from wakefulness. In terms of motor behaviour, motility decreases (depth increases) from stages 1 through 4. By criteria of sensory responsivity, thresholds generally increase (sleep deepens) from stages 1 through 4. By most physiological criteria, NREM stages 3 and 4 are particularly deactivated (deep). Thus, gradations within NREM sleep do seem fairly consistent, with a continuum extending from the "lightest" stage 1 to the "deepest" stage 4.

Problems of defining depth in REM sleep

Relative to NREM sleep, is REM sleep light or deep? The answer seems to be that by some criteria REM sleep is light and by others it is deep. For example, in terms of muscle tone, which is at its lowest point during sleep in REM sleep, it is deep. In terms of its increased rates of intermittent fine body movements, REM sleep would have to be considered light. Arousal thresholds during REM sleep are variable, apparently as a function of the meaningfulness of the stimulus (and of the possibility of its incorporation into an ongoing dream sequence). With a meaningful stimulus (*e.g.,* one that cannot be ignored with impunity), the capacity for responsivity can be demonstrated to be roughly equivalent to that of "light" NREM sleep (stages 1 and 2). With a stimulus having no particular significance to the sleeper, thresholds can be rather high. The discrepancy between these two conditions suggests an active shutting out of irrelevant stimuli during REM sleep. By most physiological criteria related to the autonomic and central nervous systems, REM sleep clearly is more like wakefulness than like NREM sleep, but drugs that cause arousal in wakefulness, such as amphetamine, sup-

press REM sleep. In terms of subjective response, recently awakened sleepers often describe REM sleep as having been "deep" and NREM sleep as having been "light." The subjectively felt depth of REM sleep may reflect the immersion of the sleeper in the vivid dream experiences of this stage.

Thus, as was true in defining sleep itself, there are difficulties in achieving unequivocal definitions of sleep depth. Several different criteria may be employed, and they are not always in agreement. REM sleep is particularly difficult to classify along any continuum of sleep depth. The current tendency is to consider it a unique state, sharing properties of both light and deep sleep. The fact that selective deprivation of REM sleep (elaborated below) results in a selective increase in such sleep on recovery nights is consistent with this view of REM sleep as unique.

Autonomic variables. Some autonomic physiological variables have a characteristic pattern relating their activity to cumulative sleep time, without respect to whether it is REM or NREM sleep. These variables are viewed by some authorities as incidental rather than essential features of the state of sleep, which is conceived in terms of the central nervous system. Such variables presumably reflect constant or slowly changing features of both kinds of sleep, such as the cumulative effects of immobility and of relaxation of skeletal muscles on metabolic processes. Body temperature, for example, drops during the early hours of sleep, reaching a low point after five or six hours, then rises toward the morning awakening.

Sleep learning, sleep-walking, sleep talking

Behavioral variables. Behaviorally, it has been shown that already established motor responses can be evoked in all stages of sleep, but it has proved much more difficult to demonstrate that new responses can be acquired during sleep. When EEG criteria of sleep are employed, it appears that "sleep learning" of verbal material takes place only to the degree that the person being tested is partially awake during the presentation of the stimuli. Another line of behavioral study is the observation of spontaneously occurring integrated behaviour patterns, such as walking and talking during sleep. In keeping with the idea of a heightened tonic (continuous) motor inhibition during REM sleep, but contrary to the idea that such behaviour is an acting out of especially vivid dream experiences or a substitute for them, sleep talking occurs primarily in NREM sleep and sleepwalking exclusively in NREM sleep. Talking in one's sleep is particularly characteristic of lighter NREM sleep (stage 2), while sleepwalking is initiated from deeper NREM sleep (stage 4). Episodes of NREM sleepwalking generally do not seem to be associated with any remembered dreams, nor is NREM sleep talking consistently associated with reported dreams of appropriate content.

For a discussion of dreaming, see below *Dreams and dreaming.*

EFFECTS OF SLEEP DEPRIVATION

One time-honoured approach to determining the function of an organ or process is to deprive an organism of the organ or process. In the case of sleep, the deprivation approach to function has been applied, both experimentally and naturalistically, to sleep as a unitary state (general sleep deprivation), and, experimentally only, to particular kinds of sleep (selective sleep deprivation).

Types of sleep deprivation

General sleep deprivation may be either total (*e.g.*, a person has had no sleep at all for a period of days) or partial (*e.g.*, over a period a person obtains only three or four hours of sleep per night). The method of general deprivation studies is enforced wakefulness. Selective deprivation has been reported for two stages of sleep: stage 4 of NREM sleep and REM sleep. Both typically occur after the appearance of other sleep stages, REM sleep after all four NREM stages and stage 4 after the lighter NREM stages. The general idea of selective deprivation studies is to allow the sleeper to have natural sleep until the point at which he enters the stage to be deprived and then to prevent the stage, either by experimental awakening or by other manipulations such as application of a mildly noxious stimulus or prior administration of a drug known to suppress it. The hope is that total sleep time will not be altered but that increased occurrence of some other stage will substitute for the loss of the one selectively eliminated.

General sleep deprivation. On a three-hour sleep schedule, partial deprivation does not reproduce, in miniaturized form, the same relative distribution of sleep patterns achieved in a seven- or eight-hour sleep period. Some increase is observed in absolute amounts of REM sleep during the three-hour sleep period as compared to the first three hours of normal sleep, and there also is a significant increase in the amount of stage 4 of NREM sleep. Lighter NREM sleep (*e.g.*, stage 2) seems to have a particularly low priority under partial sleep deprivation. Although the REM sleep percentage increases somewhat under partial deprivation, the person is still far from achieving his usual quota in absolute minutes of sleep time. On uninterrupted recovery nights following the termination of the deprivation, there is more REM sleep than there was before the deprivation. This change is viewed as a compensatory "rebound" of REM sleep such that at least some of the quota is made up. Most if not all of the nightly quota of stage 4 of NREM sleep can be achieved on a three-hour nightly schedule. Because partial deprivation on a three-hour nightly regimen also tends to be selective deprivation (the person receives most of his quota of stage 4 NREM sleep but relatively little of his quota of stage REM), the behavioral effects of such deprivation may be relevant to the question of the adaptive functions served by REM sleep. One study has reported no effects from deprivation of REM sleep on the capacity for performance on a perceptual discrimination task but decreased motivation. When a schedule of partial deprivation began to interfere with the routine accumulation of stage 4 (*i.e.*, less than three hours of sleep per night), on the other hand, the capacity for performance seemed to be adversely affected.

Effects of total sleep deprivation

In view of several obvious practical considerations, many general deprivation studies have used animals rather than human beings as experimental subjects. Waking effects routinely observed in these studies have been of deteriorated physiological functioning, sometimes including actual tissue damage. Long-term sleep deprivation in the rat (six to 33 days), accomplished by enforced locomotion of both experimental and control animals but timed to coincide with any sleep of the experimental animals, has been shown to result in severe debilitation and death of the experimental but not the control animals. This supports the view that sleep serves a vital physiological function. There is some suggestion that age is related to sensitivity to the effects of deprivation, younger organisms proving more capable of withstanding the stress than mature ones.

Among human subjects, the champion non-sleeper apparently was a 17-year-old student who voluntarily undertook a 264-hour sleep deprivation experiment. Effects noted during the deprivation period included irritability, blurred vision, slurring of speech, memory lapses, and confusion concerning his identity. No long-term (*i.e.*, post-recovery) effects were observed on either his personality or his intellect. More generally, although brief hallucinations and easily controlled episodes of bizarre behaviour have been observed after five to 10 days of continuous sleep deprivation, these symptoms do not occur in most subjects and thus offer little support to the hypothesis that sleep loss induces psychosis. In any event, these symptoms rarely persist beyond the period of sleep that follows the period of deprivation. When inappropriate behaviour does persist, it generally seems to be in persons known to have a tendency toward such behaviour. Generally, upon investigation, injury to the nervous system has not been discovered in persons who have been deprived of sleep for many days. This negative result must be understood in the context of the limited duration of these studies and should not be interpreted as indicating that sleep loss is either safe or desirable. The short-term effects observed with the student mentioned are typical and are of the sort that, in the absence of the continuous monitoring his vigil received, might well have endangered his health and safety.

Other commonly observed behavioral effects during total sleep deprivation include fatigue, inability to concentrate, and visual or tactile illusions and hallucinations. These effects generally become intensified with increased loss of

sleep, but they also wax and wane in a cyclic fashion in line with 24-hour fluctuations in EEG alpha-wave (eight to 12 hertz) phenomena and with body temperature, becoming most acute in the early morning hours. Changes in intellectual performance during moderate sleep loss can, to a certain extent, be compensated for by increased effort and motivation. In general, tasks that are work paced (the subject must respond at a particular instant of time not of his own choice) tend to be affected more adversely than tasks that are self-paced. Errors of omission are common with the former kind of task and are thought to be associated with "microsleep"—momentary lapses into sleep. Changes in body chemistry and in workings of the autonomic nervous system sometimes have been noted during deprivation, but it has proved difficult to establish either consistent patterning in such effects or whether they should be attributed to sleep loss per se or to the stress or other incidental features of the deprivation manipulation. In general, involuntary bodily functions seem relatively more impervious to effects of short-term deprivation than are adaptive, or voluntary, ones. The length of the first recovery sleep session for the student mentioned above, following his 264 hours of wakefulness, was slightly less than 15 hours. His sleep demonstrated increased amounts of both stage 4 NREM and stage REM sleep.

Effects of selective sleep deprivation

Selective sleep deprivation. Studies of selective sleep deprivation have confirmed the attribution of need for both stage 4 NREM and REM sleep, because an increasing number of experimental arousals is required each night to suppress both stage 4 and REM sleep on successive nights of deprivation, and because both show a clear rebound effect following deprivation. Rebound from stage 4 NREM-sleep deprivation occurs only on the night following termination of the deprivation regardless of the length of the deprivation, whereas the duration of the rebound effect following REM-sleep deprivation is related to the length of the prior deprivation. Little is known of the consequences of stage 4 deprivation.

Particular interest has attached to the selective deprivation of REM sleep, partly because of its unique and somewhat puzzling properties as an activated state of sleep and partly because of the association of this stage with vivid dreaming. REM-sleep-deprivation studies once were considered also to be "dream-deprivation" studies. This psychological view of REM sleep deprivation has become less pervasive since the experimental demonstration of the occurrence of dreaming during NREM-sleep stages, and because, contrary to the Freudian position that the dream is an essential safety valve for the release of emotional tensions, it has become evident that REM-sleep deprivation is not psychologically disruptive and may in fact be helpful in treating depression. REM-sleep-deprivation studies have focused more upon the presumed functions of the REM state than upon those of the vivid dreams that accompany it. The evidence from these studies has proved to be partially supportive of a number of different theoretical positions concerning REM sleep. Some animal studies have reported deleterious effects of REM-sleep deprivation on learning or other cognitive tasks (*i.e.*, tasks concerned with thinking, remembering, perceiving, and the like), in line with the view that cognitive processing may be one function of REM sleep. Other animal studies have shown heightened levels of sexuality and aggressiveness after a period of deprivation, suggesting a drive-regulative function for REM sleep. Other observations suggest increased sensitivity of the central nervous system to auditory stimuli and to electroconvulsive shock following deprivation, as might have been predicted from the theory that REM sleep somehow serves to maintain CNS integrity.

Effects of REM-sleep deprivation

Although there is a need for REM sleep, apparently it is not absolute. Animals have been deprived of REM sleep for as long as two months without showing behavioral or physiological evidence of injury. Several problems arise in connection with the methods of most REM-sleep-deprivation studies. Controls for factors such as stress, sleep interruption, and total sleep time are difficult to manage. Thus, it is unclear whether observed effects of REM-sleep deprivation are the result of REM-sleep loss or the result of such factors as stress and general sleep loss. It also is unclear whether it is the loss of continuous REM sleep or of the intermittent events that accompany it that is crucial in REM-sleep deprivation. Preliminary research indicates the latter, suggesting that REM-deprivation studies are more relevant to the function of separate intermittent events occurring in sleep than to the function of the continuous REM sleep.

PATHOLOGICAL ASPECTS

It is important, at the outset, to emphasize that, as dramatic and reliable as the various stages of sleep are, their functions or relations to waking performance, mood, or health are still largely unknown. Thus, association of a sleep abnormality with a certain stage of sleep (either in the sense that an abnormal event occurs during a certain stage or in the sense that an abnormal condition is associated with an increase or decrease in the proportion of total sleep time spent in that stage) is difficult to interpret when the function or necessity of that stage is uncertain. The pathology of sleep includes: (1) primary disturbances of sleep–wakefulness mechanisms, such as seem to characterize encephalitis lethargica ("sleeping sickness"), narcolepsy (irresistible brief episodes of sleep), and hypersomnia (sleep attacks of lesser urgency but greater duration than those of narcolepsy); (2) minor episodes occurring during sleep, such as bed-wetting and nightmares; (3) medical disorders such as sleep apnea whose symptoms occur during sleep; (4) sleep symptoms of the major psychiatric disorders; and (5) disorders of sleep schedule.

Primary disturbances. Epidemic lethargic encephalitis is produced by viral infections of sleep–wakefulness mechanisms in the hypothalamus, a structure at the upper end of the brain stem. The disease often passes through several stages: fever and delirium; hyposomnia (loss of sleep); and hypersomnia (excessive sleep), sometimes bordering on a coma. Inversions of 24-hour sleep–wakefulness patterns also are commonly observed, as are disturbances in eye movements.

Narcolepsy, like encephalitis, is thought to involve specific abnormal functioning of subcortical sleep regulatory centres. Some people who experience attacks of narcolepsy also have one or more of the following auxiliary symptoms: cataplexy, a sudden loss of muscle tone often precipitated by an emotional response such as laughter or startle and sometimes so dramatic as to cause the person to fall down; hypnagogic (sleep onset) and hypnopompic (awakening) visual hallucinations of a dreamlike sort; and hypnagogic or hypnopompic sleep paralysis, in which the person is unable to move voluntary muscles (except respiratory muscles) for a period ranging from several seconds to several minutes. When narcolepsy includes one or more of the accessory symptoms, some of the sleep attacks consist of periods of REM at onset of sleep. This precocious triggering of REM sleep (which occurs in adults generally only after 70–90 minutes of NREM sleep) may indicate that the accessory symptoms are dissociated aspects of REM sleep; *i.e.*, the cataplexy and the paralysis represent the active motor inhibition of REM sleep and the hallucinations represent the dream experience of REM sleep. Thus, narcolepsy involves REM sleep, and it is thought that it probably involves a failure of wakefulness mechanisms to inhibit the REM-sleep mechanisms.

Hypersomnia; hyposomnia

Hypersomnia may involve either excessive daytime sleep and drowsiness or a nocturnal sleep period of greater than normal duration but does not include sleep-onset REM periods. One reported concomitant of hypersomnia, the failure of heart rate to decrease during sleep, suggests that hypersomniac sleep may not be as restful per unit of time as is normal sleep. In its primary form, hypersomnia is probably hereditary in origin (as is also narcolepsy) and is thought to involve some disruption of the functioning of hypothalamic sleep centres. Narcolepsy and hypersomnia are not characterized by grossly abnormal EEG sleep patterns. The abnormality seems to involve a failure in "turn on" and "turn off" mechanisms regulating sleep, rather than in the sleep process itself. Narcoleptic and hypersomniac symptoms can be managed by administration of drugs. Several forms of hypersomnia are periodic rather than chronic. One rare disorder of periodically excessive

sleep, the Kleine-Levin syndrome, is characterized by periods of two to four weeks of excessive sleep along with a ravenous appetite and psychotic-like behaviour during the few waking hours. The "Pickwickian syndrome" (in reference to the fat boy, Joe, in Dickens' *Pickwick Papers*), another form of periodically excessive sleep, is associated with obesity and respiratory insufficiency.

Hyposomnia (this word, meaning "too little sleep," is chosen in preference to "insomnia," or "lack of sleep," because some sleep invariably is present) is less clearly understood than the conditions already mentioned. It has been demonstrated that, by physiological criteria, self-described poor sleepers generally sleep much better than they imagine. Their sleep, however, does show signs of disturbance: frequent body movement, enhanced levels of autonomic functioning, reduced levels of stage REM, and in some the intrusion of waking rhythms (alpha waves) throughout the various sleep stages. Although hyposomnia in a particular situation is common and without pathological import, chronic hyposomnia may be related to psychological disturbance. Hyposomnia conventionally is treated by administration of drugs but often with substances that are potentially addictive and otherwise dangerous when used over long periods. Newer treatments involve behavioral programs such as the temporary restriction of sleep time and its gradual reinstatement.

Minor episodes. Among the minor episodes sometimes considered abnormal in sleep are: somniloquy (sleep talking) and somnambulism (sleepwalking), enuresis (bedwetting), bruxism (tooth grinding), snoring, and nightmares. Sleep talking seems more often to consist of inarticulate mumblings than of extended, meaningful utterances. It occurs at least occasionally for many people and at this level cannot be considered pathological. Sleepwalking is not uncommon in children, but its continuation into adulthood is suggestive of persistent immaturity of the central nervous system. Enuresis may be a secondary symptom of a variety of organic conditions or, more frequently, a primary disorder in its own right. In the latter case, it seems to involve some immaturity in neural control of bladder muscles. While mainly a disorder of early childhood, enuresis persists into adulthood for a small number of persons. Treatment generally has been directed either toward sensitizing the sleeper to bladder distention, so that he will awaken and urinate according to appropriate social norms, or toward increasing bladder capacity. Primary enuresis does not seem to be an abnormality of sleep, sleep cycles of bed-wetting and of normal children being roughly the same. Tooth grinding is not consistently associated with any particular stage of sleep, nor does it appreciably affect overall sleep patterning; it, too, seems to be an abnormality in rather than of sleep.

Sleep talking; sleep-walking; bed-wetting; tooth grinding

Varieties of nightmare

A variety of frightening experiences associated with sleep, at one time or another, have been called nightmares. Because not all such phenomena have proved to be identical in their associations with sleep stages or with other variables, several distinctions need to be made among them. Incubus, the classic nightmare of adult years, consists of arousal from stage 4 NREM sleep with a sense of heaviness over the chest, with diffuse anxiety, but with little or no dream recall. Night terrors (*pavor nocturnus*) are disorders of early childhood. Delta-wave NREM sleep is suddenly interrupted with a scream; the child may sit up in apparent terror and be incoherent and inconsolable. After a period of minutes, he returns to sleep, often without ever having been fully alert or awake. Dream recall generally is absent, and the entire episode may be forgotten in the morning. Anxiety dreams most often seem associated with spontaneous arousals from REM sleep. There is remembrance of a dream whose content is in keeping with the disturbed awakening. While their persistent recurrence probably indicates waking psychological disturbance or stress caused by a difficult situation, anxiety dreams occur occasionally in many otherwise healthy persons.

Disorders accentuated during sleep. A variety of medical symptoms may be accentuated by the conditions of sleep. Attacks of angina (spasmodic, choking pain), for example, apparently can be augmented by the activation of the autonomic nervous system in REM sleep; the same is true of gastric acid secretions in persons who have duodenal ulcers. NREM sleep, on the other hand, can increase the likelihood of certain kinds of epileptic discharge.

Rhythmic snoring, which can occur throughout sleep, indicates the partial muscular relaxation of sleep, and its occasional occurrence is not abnormal. When snoring is of the loud, laboured, snorting variety, however, and is accompanied by pauses in respiration of more than 10 seconds in duration, broken by gasping sounds, the respiratory disorder called sleep apnea may be present. This disorder can occur at any age but is most common in the elderly. It results in hypoxia and sleep fragmentation, both of which contribute to excessive daytime sleepiness and cognitive deficits. Treatment approaches include behaviour change (reduction of alcohol consumption and body weight), sleep position training, mechanical appliances to keep the airway unobstructed, and surgery.

The resemblance of dream consciousness to waking psychotic experience often has been noted, and the psychotic has been considered a "waking dreamer." Thus, it has been theorized that waking psychotic symptoms may be generated by a spontaneous, or REM-sleep-deprivation-induced, shift of REM phenomena from sleep to the waking state. Symptomatically, schizophrenics have shown neither the exacerbation of psychotic symptoms under experimental REM-sleep deprivation nor the consistent or large deviations from normal EEG sleep patterning that would seem to be required by the hypothesis that sleep mechanisms play some critical role in bringing on psychotic episodes. Depressed people do sleep less and have an earlier first REM period than normal people. The first REM period, occurring 40–60 minutes after sleep onset, is often longer than normal with more eye movement activity. This suggests a disruption in the drive-regulation function, affecting such things as sexuality, appetite, or aggressiveness, all of which are reduced in such persons. REM deprivation by pharmacological agents (tricyclic antidepressants) or by REM-awakening techniques appears to reverse this sleep abnormality and to relieve the waking symptoms.

Disorders of sleep schedule. There are two prominent types of sleep-schedule disorder: phase-advanced sleep and phase-delayed sleep. In the former the sleep onset and offset occur earlier than the social norms, and in the latter sleep onset is delayed and waking is also later in the day than is desirable. These alterations in the sleep–wake cycle may occur in shift workers or following international travel across time zones. They can be treated by gradual readjustment of the timing of sleep.

DRUGS AND SLEEP

Various chemical substances long have been employed to induce or prolong sleep, but there have been few controlled, double-blind studies (neither the physician who evaluates the results nor the patient knows whether the latter has received a drug or placebo—an inert substitute) of alleged hypnotics (sleep-inducing drugs) in which sleep has been assessed by physiological measurement; and the mechanisms of sleep themselves are only now beginning to be isolated. The little research that has been done makes it clear that the manner in which a drug affects sleep can be extremely complex, with different effects sometimes attributable to different dosages of the same substance and with different effects sometimes observed for short-term and long-term administration of the same substance.

Drugs that reduce REM sleep

Many pharmacological agents tend to reduce the absolute amount and relative proportion of sleep spent in REM sleep. In this sense, REM sleep has been called a fragile state. Specifically, most effective hypnotics, particularly the barbiturates (*e.g.*, pentobarbital, secobarbital), decrease both total REM time and the proportion of sleep spent in REM sleep, with enhanced amounts of NREM sleep. Amphetamine, an analeptic (stimulant), decreases REM sleep. Many tranquilizers also slightly reduce REM sleep. There is evidence that the withdrawal symptoms of persons taken off addictive drugs of any variety (*e.g.*, barbiturates, amphetamines, narcotics) are accompanied by relatively high percentages of REM sleep. It has been suggested that the drugs in question are REM-sleep deprivers, that the elevated periods spent in REM sleep on with-

drawal represent REM-sleep rebound, that the withdrawal syndrome may be functionally related to high pressure for REM sleep, and that the vivid, unpleasant dreams associated with REM-sleep rebound may be responsible for some patients' return to the use of the REM-sleep-depriving agents. Caffeine seems to have little effect on normal sleep patterning, but the effects of alcohol are variable: the short-term effect is to reduce the time spent in REM sleep, but, with continued use, there may be a REM-sleep rebound. Not all drug effects are on REM sleep; some of the more recently developed tranquilizers and hypnotics have been found to reduce stage 4 of NREM sleep.

Much interest has attached to the search for hypnotic substances that are not REM-sleep deprivers, that is, that induce or prolong sleep without altering natural sleep patterns. While some such hypnotics have been found, they most often either have adverse side effects or have not been fully evaluated. Theoretically, the most interesting substances are those few that have been found to increase REM sleep. In certain dosage ranges and under certain conditions, such an effect has been noted for reserpine, a tranquilizer, and for D-lysergic acid diethylamide (LSD), a hallucinogen. Both substances have important interactions with neurohumours (serotonin and norepinephrine—substances formed in nerve cells), and their effects may offer clues to the mechanisms underlying REM sleep.

THEORIES OF SLEEP

Two kinds of sleep theory of contemporary interest may be distinguished. One begins with the peripheral physiology of sleep and relates it to underlying neural (nervous system) or biochemical mechanisms. Such theories most often rely on experiments with animals by means of drugs or surgery. Alternatively, sleep theories may start with behavioral observations of sleep and may attempt to specify the functions of such a state of lethargy and insensitivity from an evolutionary or adaptive point of view. The question here is not so much how people sleep, or even why they sleep, but what good it does.

Mechanistic theories. Historically, mechanistic theories of sleep have focused on a succession of organs or structures in a manner reflective of the degree of access different civilizations have had to the inner workings of the human body. Thus, the relatively perceptible processes of circulation, digestion, and secretion played large roles in the theories of classical antiquity, and modern theories have been concerned with the central nervous system, particularly the brain, although various peripheral factors in the induction of sleep are not ruled out. Proposals that blood composition, metabolic changes, or internal secretions regulate sleep are necessarily incomplete to the extent that they ignore the contributions of environment and intent to the onset of sleep. It also has been noted that in two-headed human monsters one "twin" may seem asleep while the other is awake, despite their sharing a circulatory system.

Neural theories. Among neural theories of sleep, there are certain issues that each must face. Is the sleep–wakefulness alternation to be considered a property of individual neurons (nerve cells), making unnecessary the postulation of specific regulative centres, or is it to be assumed that there are some aggregations of neurons that play a dominant role in sleep induction and maintenance? The Russian physiologist Ivan Petrovich Pavlov adopted the former position, proposing that sleep is the result of irradiating inhibition among cortical and subcortical neurons (nerve cells in the outer brain layer and in the brain layers beneath the cortex). Microelectrode studies, on the other hand, have revealed high rates of discharge during sleep from many neurons in the motor and visual areas of the cortex, and thus it seems that, as compared with wakefulness, sleep must consist of a different organization of cortical activity rather than some general, overall decline.

Pavlov's theory

Another issue has been whether there is a waking centre, fluctuations in whose level of functioning are responsible for various degrees of wakefulness and sleep, or whether the induction of sleep requires another centre, actively antagonistic to the waking centre. Early speculation favoured the passive view of sleep. A *cerveau isolé* preparation, an animal in which a surgical incision high in the midbrain has separated the cerebral hemispheres from sensory input, demonstrated chronic somnolence. It has been reasoned that a similar cutting off of sensory input, functional rather than structural, must characterize natural states of sleep. Other supporting observations for the stimulus-deficiency theory of sleep included presleep rituals, such as turning out the lights, regulating stimulus input, and the facilitation of sleep induction by muscular relaxation. With the discovery of the ascending reticular activating system (ARAS, a network of nerves in the brain stem), it was found that it is not the sensory nerves themselves that maintain cortical arousal but rather the ARAS, which projects impulses diffusely to the cortex from the brain stem. Presumably sleep would result from interference with the active functioning of the ARAS. Injuries to the ARAS were, in fact, found to produce sleep. Sleep thus seemed passive, in the sense that it was the absence of something (ARAS support of sensory impulses) characteristic of wakefulness.

Theory has tended to depart from this belief and to move toward conceiving of sleep as an actively produced state. Two kinds of observation primarily have been responsible for the shift. First, earlier studies showing that sleep can be induced directly by electrical stimulation of certain areas in the hypothalamus have been confirmed and extended to other areas in the brain. Second, the discovery of REM sleep has been even more significant in leading theorists to consider the possibility of actively produced sleep. REM sleep, by its very active nature, defies description as a passive state. As is noted below, REM sleep can be eliminated in experimental animals by the surgical destruction of a group of nerve cells in the pons, the active function of which appears to be necessary for REM sleep. Thus, it is difficult to imagine that the various manifestations of REM sleep reflect merely the deactivation of wakefulness mechanisms.

The REM–NREM-sleep dichotomy poses a third issue for the theories of sleep mechanisms, or at least for those that accept the idea of sleep as an active phenomenon. Does one hypnogenic (sleep-causing) system serve both kinds of sleep, or are there two antagonistic sleep systems, one for REM sleep and one for NREM sleep? Opinion is sharply divided. One group of theorists states that there must be two sleep systems. It is noted that NREM sleep is not affected, but that REM sleep is abolished, by injuries to the pontine tegmentum (the posterior part of the pons) and that NREM sleep is suppressed in animals whose brain stem has been severed at the midpoint of the pons, suggesting that an NREM-sleep centre behind this section no longer is capable of suppressing the effect of the ARAS. It is further observed that the neurohumour serotonin is localized in the brain-stem regions presumed to be responsible for NREM sleep; that destruction of serotonin-containing nerve cells in the brain stem may produce insomnia; that, in some species, reductions of serotonin by chemical interference with its production produces an amount of sleep loss correlated with the reduction of serotonin; that administration of a serotonin precursor (a substance from which serotonin is formed) after interference with production of serotonin produces a sleeplike state and that artificially induced increases in brain serotonin increase NREM sleep; that the neurohumour norepinephrine is localized in the brain-stem regions presumed to be responsible for REM sleep; and that substances interfering with the synthesis of norepinephrine suppress REM sleep. Other theorists have proposed that REM and NREM sleep are served by a common hypnogenic system. Chemical stimulation of certain brain structures, assumed to constitute a hypnogenic system, has been found capable of inducing both stages of sleep. It also is argued that different varieties of sleep should require different mechanisms no more than do different varieties of wakefulness (*e.g.*, alertness, relaxation).

One- and two-system theories

Functional theories. Functional theories stress the recuperative and adaptive value of sleep. Sleep arises most unequivocally in animals that maintain a constant body temperature and that can be active at a wide range of environmental temperatures. In such forms, increased

metabolic requirements may find partial compensation in periodic decreases in body temperature and metabolic rate (*i.e.*, during NREM sleep). Thus, the parallel evolution of temperature regulation and NREM sleep has suggested to some authorities that NREM sleep may best be viewed as a regulatory mechanism conserving energy expenditure in species whose metabolic requirements are otherwise high. As a solution to the problem of susceptibility to predation that comes with the torpor of sleep, it has been suggested that the periodic reactivation of the organism during sleep better prepares it for fight or flight, and that the possibility of enhanced processing of significant environmental stimuli during REM sleep may even reduce the need for sudden confrontation with danger. Other functional theorists agree that NREM sleep may be a state of "bodily repair," while suggesting that REM sleep is one of "brain repair" or restitution, a period, for example, of increased cerebral protein synthesis or of "reprogramming" the brain so that information achieved in wakeful functioning is most efficiently assimilated. In their specification of functions and provision of evidence for such functions, such theories are necessarily vague and incomplete. The function of stage 2 NREM sleep is still unclear, for example. Such sleep is present in only rudimentary form in subprimate species yet consumes approximately half of human sleep time. Comparative, physiological, and experimental evidence is unavailable to suggest why so much human sleep is spent in this stage. In fact, poor sleepers whose laboratory sleep records show high proportions of stage 2 and little or no REM sleep often report feeling they have not slept at all.

(D.F./R.D.C.)

Dreams and dreaming

DIVERSE VIEWS ON THE NATURE OF DREAMS

Dreams as reflecting reality. Philosophers continue to argue about the differences between reality and dreams. The English philosopher Bertrand Russell (1872–1970) wrote, "It is obviously possible that what we call waking life may be only an unusual and persistent nightmare," and he further stated that "I do not believe that I am now dreaming but I cannot prove I am not." Philosophers generally try to resolve the question by saying that so-called waking experience seems vivid and coherent. As the French philosopher René Descartes (1596–1650) put it: " . . . memory can never connect our dreams one with the other or with the whole course of our lives as it unites events which happen to us while we are awake"; or, as Russell stated succinctly, "Certain uniformities are observed in waking life, while dreams seem quite erratic."

Dreams in various cultures

Members of many cultures have variously coped with this dilemma; for example, among the Eskimo of Hudson Bay and the Patani Malay people, it is believed that during sleep one's "soul" leaves the body to live in a special dreamworld. Believers often consider it dangerous to wake someone lest his "soul" be lost. On these grounds the Tajal people of Luzon, for example, severely punish for awakening a sleeping person. In other cultures, dream events are held to be identical with reality; thus, a Macusi Indian of Guyana is reported to have become enraged at the European leader of an expedition when he dreamed that the leader had made him haul a canoe up dangerous cataracts. He awoke exhausted and could not be persuaded that the dream was not real. There is a tradition in Borneo that if a man dreams that his wife is an adulteress, her father must take her back. A Zulu man is said to have broken off a friendship after dreaming that the friend meant him harm. A Paraguayan Indian, reportedly having dreamed that a missionary shot at him, attempted to kill the missionary.

In other instances, dream events may be believed to demand fulfillment. Jesuit priests in the 1700s reported that among Iroquois Indians it was obligatory to carry out dreams as soon as possible; one Indian was said to have dreamed that 10 friends dove into a hole in the ice of a lake and came up through another. When told of the dream, the friends duly enacted their roles in it, but unfortunately, only nine of them succeeded. After dreaming of something valuable, Kurdish people were traditionally expected to take it, by force if necessary. Among some natives of Kamchatka a man need only dream of a girl's favour for her to owe him her sexual favours.

Such interpretations in which the dream is given a status of reality need not imply that the two are indistinguishable. In some instances, the dream may be differentiated from reality, but dreams are accorded a superior status to the banal activities of wakefulness.

Dreams as a source of divination. There is an ancient belief that dreams predict the future; the Chester Beatty Papyrus is a record of Egyptian dream interpretations dating from the 12th dynasty (1991–1786 BC). In the *Iliad,* Agamemnon is visited in dream by a messenger of the god Zeus to prescribe his future actions. From India, a document called the Atharvaveda, attributed to the 5th century BC, contains a chapter on dream omens. A Babylonian dream guide was discovered in the ruins of the city of Nineveh among tablets from the library of the emperor Ashurbanipal (668–627 BC). The Old Testament is rife with prophetic dreams, those of the pharaohs and of Joseph and Jacob being particularly striking. Among pre-Islāmic peoples dream divination so heavily influenced daily life that the practice was formally forbidden by Muḥammad (*c.* 570?–632), founder of the Muslim religion.

Dreams in the Bible

Ancient and religious literatures express the most confidence about so-called message dreams. Characteristically, a god or some other respected figure appears to the dreamer (typically a king, a hero, or a priest) in time of crisis and states a message. Such reports are found on ancient Sumerian and Egyptian monuments; frequent examples appear in the Bible. Joseph Smith (1805–44), the founder of Mormonism, said that an angel had directed him to the location of buried golden tablets that described American Indians as descendants of the tribes of Israel.

Not all dream prophecies are so readily accepted. In the *Odyssey,* for example, dreams are classed as false ("passing through the Gate of Ivory") and as true ("passing the Gate of Horn"). Furthermore, prophetic meaning may be attributed to dream symbolism. In the Bible, Joseph interpreted sheaves of grain and the Moon and stars as symbols of himself and of his brethren. In general, the social status of dream interpreters varies; in cultures for which dreams loom important, their interpretation frequently is an occupation of priests, elders, or medicine men.

Perhaps the most famous dream interpretation book is that of the Greek soothsayer Artemidorus Daldianus (*c.* 2nd century AD), the *Oneirocritica* (from the Greek *oneiros,* "a dream"). Dream books remain widely available today. They continue to enjoy profitable sales everywhere among people who follow them in affairs of the heart, in gambling, and in matters of health and work.

Dreams as curative. So-called prophetic dreams in the Middle Eastern cultures of antiquity often were combined with other means of prophecy, such as animal sacrifice, and with efforts to heal the sick. In classical Greece, dreams became directly associated with healing; ailing people came to dream in oracular temples where priests and priestesses advised about the cures dreams were held to provide. Similar practices, known as dream incubation, are recorded for Babylon and Egypt. In a widespread cult, suffering petitioners came to at least 600 temples of the Greek god of medicine to perform rites or sacrifices in efforts to dream appropriately, sleeping in wait of the appearance of the god or his emissary to deliver a cure. Many stone monuments placed at the entrances of the temples survive to record dream cures.

Dreams as extensions of the waking state. Even in early human history dreams also were interpreted as reflections of waking experiences and of emotional needs. Aristotle (384–322 BC), despite his contemporaries who practiced divination and incubation, in his work *Parva Naturalia* (*On the Senses and Their Objects*) attributed dreams to sensory impressions from "external objects . . . pauses within the body . . . eddies . . . of sensory movement often remaining like they were when they first started, but often too broken into other forms by collision with obstacles." In anticipation of psychoanalyst Sigmund Freud (1856–1939), Aristotle wrote that sensory function is reduced in sleep, favouring the susceptibility of dreams to emotional subjective distortions.

In spite of Aristotle's unusually modern views and even after a devastating attack by the Roman statesman Marcus Tullius Cicero (106–43 BC) on dream divination (*De divinatione;* "On Divination"), views that dreams have supernatural attributes persisted vigorously until the 1850s and the classic work of the French physician Alfred Maury, who studied more than 3,000 reported recollections of dreams. Maury concluded that dreams arose from external stimuli, instantaneously accompanying such impressions as they acted upon the sleeping person. He wrote that part of his bed once fell on the back of his neck and woke him, leaving the memory of dreaming that he had been brought before a French revolutionary tribunal, questioned, condemned, led to the scaffold, and bound by the executioner, and that the guillotine blade had fallen.

Creative dreaming

The English poet Samuel Taylor Coleridge reported that he had written "Kubla Khan" as the result of creative thinking in a dream. Having fallen asleep while reading about that Mongol conqueror, he woke to write down a fully developed poem he seemed to have composed while dreaming. Novelist Robert Louis Stevenson said that much of his writing was developed by "little people" in his dreams, and specifically cited the story of Dr. Jekyll and Mr. Hyde in this context. The German chemist F.A. Kekulé von Stradonitz attributed his interpretation of the ring structure of the benzene molecule to his dream of a snake with its tail in its mouth. Otto Loewi, the German physiologist, attributed to a dream inspiration for an experiment with a frog's nerve that helped him win the Nobel Prize. In all of these cases the dreamers reported having thought about the same topics over considerable periods while they were awake.

Psychoanalytic interpretations. Among Freud's earliest writings was *The Interpretation of Dreams* (1899). His insistence that dreams are "the royal road to the unconscious" continued from it down to his last published statement on dreams, printed about a year before he died. Freud held that dreams reflect waking experience; he offered a theoretical explanation for their bizarre nature, invented a system for their interpretation, and elaborated on their curative potential.

Freud theorized that thinking during sleep tends to be primitive and regressive and that the effects of forgetting (repression) are reduced. Repressed wishes, particularly those associated with sex and hostility, were said to be released in dreams when the inhibitory demands of wakefulness diminished. The content of the dream was said to derive from such stimuli as urinary pressure in the bladder, traces of experiences from the previous day (day residues), and associated infantile memories. The specific dream details were called their manifest content; the presumably repressed wishes being expressed were called the latent content. Freud suggested that the dreamer kept himself from waking and avoided unpleasant awareness of repressed wishes by disguising them as bizarre manifest content in an effort called dreamwork. He held that impulses one fails to satisfy when awake are expressed in dreams as sensory images and scenes. In dreaming, Freud believed:

Dreamwork

> All of the linguistic instruments . . . of subtle thought are dropped . . . and abstract terms are taken back to the concrete . . . . The copious employment of symbols . . . for representing certain objects and processes is in harmony (with) the regression of the mental apparatus and the demands of censorship.

Freud theorized that one aspect of manifest content could come to represent a number of latent elements (and vice versa) through a process called condensation. Further displacement of emotional attitudes toward one object or person theoretically could be displaced in dreaming to another object or person or not appear in the dream at all. Freud further observed a process called secondary elaboration, which occurs when people wake and try to remember dreams. They may recall inaccurately in a process of elaboration and rationalization and provide "the dream, a smooth facade, (or by omission) display rents and cracks." This waking activity he called secondary revision.

In seeking the latent meaning of a dream, Freud advised the individual to associate freely about it. From listening to the associations, the analyst was supposed to determine what the dream represented, in part through an understanding of the personal needs of the dreamer.

Carl Jung (1875–1961) disagreed with Freud's view of dreams as being complementary to waking mental life with respect to specific instinctual impulses. Jung felt that dreams are instead compensatory, that they balance whatever elements of character are underrepresented in the way people are living their lives. Dreaming, to Jung, represents a continuous 24-hour flow of mental activity that surfaces in sleep when conditions are right, but which affects waking life when a person's behaviour denies important elements of his true personality.

Thus, dreams are constructed not to conceal or disguise forbidden wishes but to bring the under-attended areas to attention. This function is carried out unconsciously in sleep when people are living well-balanced lives. If this is not the case there may be first bad moods, then symptoms in waking. Then and only then do dreams need to be interpreted. This is best done not with a single dream and multiple free associations but with a series of dreams so that the repetitive elements become apparent.

EFFORTS TO STUDY DREAMING

Dream reports. Though each person seems to know his own private dreams, the manner in which people dream obviously defies direct observation. It has been said that each dream "is a personal document, a letter to oneself" and must be inferred from the observable behaviour of people. Furthermore, observational methods and purposes clearly affect conclusions to be drawn about the inferred dreams. Reports of dreams collected from people after morning awakenings at home tend to exhibit more content of an overt sexual and emotional nature than do those from laboratory subjects. Such experiences as dreaming in colour seldom are spontaneously mentioned but often emerge under careful questioning. Reports of morning dreams are typically richer and more complex than those collected early at night. Immediate recall differs from what is reported after longer periods of wakefulness; psychoanalysts seem to elicit more recollections of overt sexual dreams than do laboratory investigators. In spite of these complications, there have been substantial efforts to describe the general characteristics of what people say they have dreamed.

The reported length of dreams varies widely between and within individuals (and by inference, so does the length of the presumed dreams themselves). Spontaneously reported dreams among laboratory subjects are typically short; about 90 percent of these reports are less than 150 words long, although some may exceed 1,000. With additional probing, about a third of such reports are longer than 300 words.

Length of dream reports

Some investigators have been surprised by repeated findings that suggest dreams may be less fantastic or bizarre than generally supposed. In the language of modern art, one investigator stated that visual dreams are typically faithful to reality (representational) with little, if any, abstractionist or surrealistic dreaming. Any variations from the representational were characterized as impressionistic. Except for those that are very short, dreams are reported to take place in ordinary physical settings, about half of them seeming quite familiar to the dreamer; only rarely is the setting said to be exotic or peculiar.

Apparently dreams are quite egocentric, the dreamer perceiving himself as a participant, though the presence of others is typically recalled. Seldom does the person remember an empty, unpopulated dreamworld, and individuals seem to dream roughly two-thirds of the time about people they know. Usually they are close acquaintances; family members are mentioned in about 20 percent of dream reports. Recollections of notables or weird representations of people are generally rare.

The typical report is of visual imagery; indeed, in its absence, the person may say only that he had been thinking rather than "dreaming" while asleep. Rare statements about dreams dominated by auditory experience commonly are made with claims of actually having been awake. It is unusual, however, to hear of dreams with-

out some auditory characteristics. One typically is told of bland dreams; when there are emotional overtones, they tend to be unpleasant about two-thirds of the time. Fear and anxiety are most commonly mentioned, followed by anger; pleasant feelings are most often those of friendliness. Reports of overtly erotic dreams, particularly among those gathered in laboratories, are infrequent.

Despite their generally representational nature, dreams seem somehow odd or strange. Perhaps this is related to discontinuities in time and purpose. One suddenly may dream of himself in a familiar auditorium viewing a fencing match rather than hearing a lecture and abruptly in the "next scene" walking beside a swimming pool. Or a person may have the experience of lying in a hallway listening to two people standing by an elevator; he may be looking at a bleeding hand and walk across an empty room to a liquor cabinet to find a roll of adhesive tape. These sudden transitions contribute to the dreamer's feeling of strangeness, and this is enhanced by his waking statements that the bulk of his dreams cannot be clearly recalled, giving them a dim, mysterious quality.

Physiological dream research. A new era of dream research began in 1953 with the discovery that rapid eye movements during sleep seem often to signal that a person is dreaming. In that year it was observed that, about an hour or so after falling asleep, laboratory subjects are apt to experience a burst of rapid eye movement (REM) under their closed lids, accompanied by a change in brain waves detected by an electroencephalograph as an electrical pattern resembling that of an alert waking person (see above *Rapid eye movement sleep*). When subjects were awakened during REM, they reported vivid dreams 20 out of 27 times; when roused during non-REM sleep, they recalled dreams in only four of 23 instances. Subsequent systematic study confirmed this relationship between REM, activated brain waves (EEG), and dream recall. Several thousand experimental studies utilizing these observable indexes of dreaming have since been conducted.

Rapid eye movement

A major finding is that the usual report of a vivid, visual dream is primarily associated with REM and activated EEG. On being aroused while exhibiting these signs, people recall dreams with visual imagery about 80 percent of the time. When awakened in the absence of them, however, people still report some kind of dream activity, though only about 30 to 50 percent of the time; in such cases they are apt to remember their sleep experiences as being relatively "thoughtlike" and realistic and as resembling the experiences of wakefulness.

Dreaming in animals

D-state (desynchronized or dreaming) sleep has been reported for all mammals studied; it has been observed, for example, among monkeys, dogs, cats, rats, elephants, shrews, and opossums; these signs also have been reported in some birds and reptiles.

Surgical destruction of selected brain structures among laboratory animals has clearly demonstrated that the D-state depends on an area within the brain stem known as the pontine tegmentum. Evidence indicates that D-state sleep is associated with a mechanism involving a bodily chemical called norepinephrine; other stages of sleep seem to involve another chemical (serotonin) in the brain. Among other physiological changes found intimately related to D-state sleep are increased variability in breathing and heart rate, relaxation of skeletal muscles (in lower animals), and, in humans, reduction of electrical activity in muscles near the base of the tongue and penile erections or increase in vaginal blood flow and uterine contractions.

When people are chronically deprived of the opportunity to manifest D-state activity (by awakening them whenever there is EEG evidence of dreaming), it appears increasingly difficult to prevent them from dreaming. On recovery nights (after such deprivation) when the subject can sleep without interruption, there is a substantial increase in the number of reports of dreaming. This rebound effect continues in some degree on subsequent recovery nights, depending on how badly the person has been deprived.

Dream recall

During D-states in the last 6½ to 7½ hours of sleep people are likely to wake by themselves about 40 percent of the time. This figure is about the same as that for dream recall, people saying they had a dream the previous night about 35 percent of the time (roughly once every three or four nights). Evidence concerning the amount and kind of dreaming also depends on how rapidly one is roused and on the intensity of his effort to recall. Some people recall dreams more often than the average, while others rarely report them. While these two groups of people show little difference in amount of D-state sleep, evidence suggests that non-recall reflects a general tendency on the part of the individual to repress or to deny his experiences.

The psychoanalytic literature is rich with reports indicating that what one dreams about reflects his needs and his immediate and remote past experience. Nevertheless, when someone in D-state sleep is stimulated (*e.g.*, by spoken word or by drops of water on his skin), the chances that he will say he has dreamed about the stimulus, or anything similar, are quite low. Studies in which people have watched vivid movies before falling asleep also indicate some possibilities of influencing dreams but again clearly emphasize the limitations of such influences. Highly suggestible people seem likely to dream as they are told to do while under hypnosis, but the influence of direct suggestion during ordinary wakefulness seems quite limited.

Variations within the usual range of about 18 to 30 percent of D-state sleep apparently are unrelated to differences in the amount or content of dreaming. The amount of D-state further seems generally independent of wide variations in the daily activities or personality characteristics of different people; groups of scientists, athletes, housewives, and artists, for example, cannot be reliably distinguished from one another in terms of D-state activity. Such disorders as schizophrenia and mental retardation appear to have no clearly discernible effect on the amount of time sufferers spend in REM-activated EEG sleep.

DREAMLIKE ACTIVITIES

Related states of awareness may be distinguished from the dream experiences typically reported; these include dreamlike states experienced as a person falls asleep and as he awakens, respectively called hypnagogic and hypnopompic reveries. During sleep itself there are nightmares, observable signs of sexual activity (*e.g.*, nocturnal emissions of sperm), and sleepwalking. Even people who ostensibly are awake may show evidence of such related phenomena as hallucinating, trance behaviour, and reactions to drugs.

Rapid eye movement is not characteristic of sleep onset; nevertheless, as people drift (as inferred from EEG activity) from wakefulness through drowsiness into sleep, they report dreamlike hypnagogic experiences about 90 percent of the time on being awakened. Most of these experiences (about 80 percent) are said to be visual. If dreaming is defined as at least partly hallucinatory and somewhat dramatic, then awakening from drowsiness or at the onset of sleep yields recall of experiences that may be classified as dreams for about 75 percent of the occasions. These "dreamlets" seem to differ from dream-associated REM sleep in being less emotional (neither pleasant nor unpleasant), more transient, and less elaborate. Such hypnagogic experiences apparently tend to incorporate abstract thinking and recall of recent events (day residues) and to be quite typical of falling asleep. Systematic studies remain to be made of the hypnopompic reveries commonly reported mornings before full arousal, but it seems likely that they include recollections of the night's dreams, or represent one's drifting back into transient REM sleep.

Extreme behavioral manifestations during sleep—night terrors, nightmares, sleepwalking, enuresis—all have been found generally unrelated to ordinary dreaming. Night terrors are characterized by abrupt awakening, sometimes with a scream; a sleeping child may sit up in bed, apparently terror-stricken, with wide-open eyes, and often with frozen posturing that may last several minutes. Afterward there typically is no recollection of dreamlike experience. Observed in about 2 or 3 percent of children, roughly half of the attacks of night terror occur between the ages of four and seven; about 10 percent of them are seen among youngsters as old as 12 to 14 years. Nightmares typically seem to be followed by awakening with feelings of suffocation and helplessness and expressions of fearful or threatening thoughts. Evidence of nightmares is ob-

Nightmares

served for 5 to 10 percent of children, primarily about eight to 10 years of age. Studies have suggested that signs of spontaneously generated night terrors and nightmares may be related to abrupt awakening from deep sleep that experimentally appears dreamless. This suggests that the vividly reported fears well may be produced by emotional disturbances that first occur on awakening.

Sleepwalking, observed in about 1 percent of children, predominantly appears between ages 11 and 14. Apparently sleeping individuals rise and walk from their beds, eyes open, usually avoiding obstacles, and expressing no recollection of the episode when they wake. Studies of EEG data indicate that sleepwalking occurs only in deep sleep when dreams seem essentially absent; the behaviour remains to be reported for REM sleep. Enuresis occurs in about one-fourth of children over age four. These episodes seem not to be associated with REM as much as they do with deep sleep in the absence of D-state signs.

Nocturnal emission of sperm remains to be described in terms of any distinguishing EEG pattern; such events are quite rarely observed among sleeping laboratory subjects. Among a large sample of males who were interviewed about their sexual behaviour about 85 percent reported having experienced emissions at some time in their lives, typical frequency during the teens and 20s being about once a month. Of the females interviewed 37 percent reported erotic dreams, sometimes with orgasm, averaging about three to four times a year. Most often, however, openly sexual dreams are said not to be accompanied by orgasm in either sex. Males not infrequently could recall no dreams associated with emission, although most implicated erotic dreaming.

Dreamlike experiences induced as trances, deliriums, or hallucinatory behaviour by drugs seem attributable to lowered efficiency of the central nervous system in processing sensory stimuli from the external environment. The result seems to be that one's physiological activities begin to escape environmental constraint to the point that internalized, uncritical thinking and perceiving prevail.

Since antiquity, dreams have been viewed as a source of divination, as a form of reality, as a curative force, and as an extension or adjunct of the waking state. Psychoanalytic theorists stress the individual meaningfulness of dreams and their relation to personal hopes and fears. Contemporary research focuses on efforts to discover and describe unique, complex biochemical and neurophysiological bases of dreaming. Among the plethora of theories ranging from those that assert dreaming to be awareness of a god's voice to those that reduce the dream to physical activity in the nervous system, no single, encompassing theory seems to be available. (W.B.W./R.D.C.)

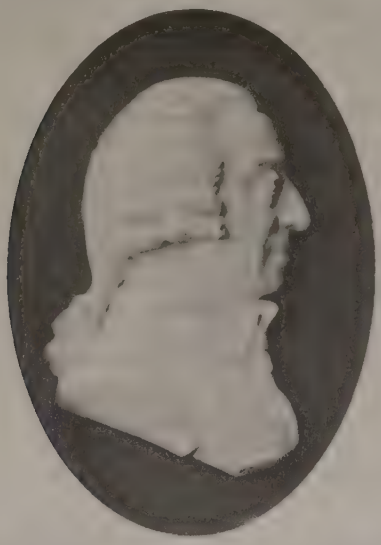
BIBLIOGRAPHY

Sleep: The standard reference on the physiology of sleep is NATHANIEL KLEITMAN, *Sleep and Wakefulness*, rev. and enl. ed. (1963). Also see ERNEST HARTMANN, *The Functions of Sleep* (1973), an excellent summary of psychological and biological research on REM and NREM sleep functions; ANTHONY KALES and JOYCE D. KALES, *Evaluation and Treatment of Insomnia* (1984), the most comprehensive work on this sleep disorder; and JERROLD S. MAXMEN, *A Good Night's Sleep: A Step-By-Step Program for Overcoming Insomnia and Other Sleep Problems* (1981), a popular book that covers a range of sleep problems.

Dreams and dreaming: Two differing classic theories of dream interpretation are found in SIGMUND FREUD, *The Interpretation of Dreams*, vol. 4 and 5 in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, edited by JAMES STRACHEY (1953, reprinted 1981; originally published in German, 8th ed., 1930), also available in other translations; and C.G. JUNG, *Dream Analysis: Notes of the Seminar Given in 1928–1930* (1984). Also see G.E. VON GRUNEBAUM and ROGER CAILLOIS (eds.), *The Dream and Human Societies* (1966), a scholarly work, with a chapter on dream research; RICHARD M. JONES, *The New Psychology of Dreaming* (1970, reissued 1978), which attempts to coordinate experimental findings on dreams with classic theories; ROSALIND DYMOND CARTWRIGHT, *Night Life: Explorations in Dreaming* (1977), which covers a series of studies approaching a laboratory-based understanding of dreams and dreaming; MONTAGUE ULLMAN and NAN ZIMMERMAN, *Working with Dreams* (1979, reissued 1985), which helps the general reader understand and work with dream material; and ERNEST HARTMANN, *The Nightmare: The Psychology and Biology of Terrifying Dreams* (1984), a comprehensive work. (D.F./W.B.W./R.D.C.)

Adam Smith

After two centuries, Adam Smith remains a towering figure in the history of economic thought. Known primarily for a single work, *An Inquiry into the nature and causes of the Wealth of Nations* (1776), the first comprehensive system of political economy, Smith is more properly regarded as a social philosopher whose economic writings constitute only the capstone to an overarching view of political and social evolution. If his masterwork is viewed in relation to his earlier lectures on moral philosophy and government, as well as to allusions in *The Theory of Moral Sentiments* (1759) to a work he hoped to write on "the general principles of law and government, and of the different revolutions they have undergone in the different ages and periods of society," then *The Wealth of Nations* may be seen not merely as a treatise on economics but as a partial exposition of a much larger scheme of historical evolution.



Smith, paste medallion by James Tassie, 1787. In the Scottish National Portrait Gallery, Edinburgh.

By courtesy of the Scottish National Portrait Gallery, Edinburgh

Early life. Unfortunately, much more is known about Smith's thought than about his life. Though the exact date of his birth is unknown, he was baptized on June 5, 1723, in Kirkcaldy, a small (population 1,500) but thriving fishing village near Edinburgh, the son by second marriage of Adam Smith, comptroller of customs at Kirkcaldy, and Margaret Douglas, daughter of a substantial landowner. Of Smith's childhood nothing is known other than that he received his elementary schooling in Kirkcaldy and that at the age of four years he was said to have been carried off by gypsies. Pursuit was mounted, and young Adam was abandoned by his captors. "He would have made, I fear, a poor gypsy," commented his principal biographer.

Influence of Hutcheson

At the age of 14, in 1737, Smith entered the university of Glasgow, already remarkable as a centre of what was to become known as the Scottish Enlightenment. There, he was deeply influenced by Francis Hutcheson, a famous professor of moral philosophy from whose economic and philosophical views he was later to diverge but whose magnetic character seems to have been a main shaping force in Smith's development. Graduating in 1740, Smith won a scholarship (the Snell Exhibition) and traveled on horseback to Oxford, where he stayed at Balliol College. Compared to the stimulating atmosphere of Glasgow, Oxford was an educational desert. His years there were spent largely in self-education, from which Smith obtained a firm grasp of both classical and contemporary philosophy.

Returning to his home after an absence of six years, Smith cast about for suitable employment. The connections of his mother's family, together with the support of the jurist and philosopher Lord Henry Kames, resulted in an opportunity to give a series of public lectures in Edinburgh—a form of education then much in vogue in the prevailing spirit of "improvement."

The lectures, which ranged over a wide variety of subjects from rhetoric to history and economics, made a deep impression on some of Smith's notable contemporaries. They also had a marked influence on Smith's own career, for in 1751, at the age of 27, he was appointed professor of logic at Glasgow, from which post he transferred in 1752 to the more remunerative professorship of moral philosophy, a subject that embraced the related fields of natural theology, ethics, jurisprudence, and political economy.

Glasgow. Smith then entered upon a period of extraordinary creativity, combined with a social and intellectual life that he afterward described as "by far the happiest, and most honourable period of my life." During the week he lectured daily from 7:30 to 8:30 AM and again thrice weekly from 11 AM to noon, to classes of up to 90 students, aged 14 to 16. (Although his lectures were presented in English, following the precedent of Hutcheson, rather than in Latin, the level of sophistication for so young an audience today strikes one as extraordinarily demanding.) Afternoons were occupied with university affairs in which Smith played an active role, being elected dean of faculty in 1758; his evenings were spent in the stimulating company of Glasgow society.

His friends and acquaintances

Among his wide circle of acquaintances were not only members of the aristocracy, many connected with the government, but also a range of intellectual and scientific figures that included Joseph Black, a pioneer in the field of chemistry, James Watt, later of steam-engine fame, Robert Foulis, a distinguished printer and publisher and subsequent founder of the first British Academy of Design, and, not least, the philosopher David Hume, a lifelong friend whom Smith had met in Edinburgh. Smith was also introduced during these years to the company of the great merchants who were carrying on the colonial trade that had opened to Scotland following its union with England in 1707. One of them, Andrew Cochrane, had been a provost of Glasgow and had founded the famous Political Economy Club. From Cochrane and his fellow merchants Smith undoubtedly acquired the detailed information concerning trade and business that was to give such a sense of the real world to *The Wealth of Nations*.

The Theory of Moral Sentiments. In 1759 Smith published his first work, *The Theory of Moral Sentiments*. Didactic, exhortative, and analytic by turns, the *Theory* lays the psychological foundation on which *The Wealth of Nations* was later to be built. In it Smith described the principles of "human nature," which, together with Hume and the other leading philosophers of his time, he took as a universal and unchanging datum from which social institutions, as well as social behaviour, could be deduced.

One question in particular interested Smith in *The Theory of Moral Sentiments*. This was a problem that had attracted Smith's teacher Hutcheson and a number of Scottish philosophers before him. The question was the source of the ability to form moral judgments, including judgments on one's own behaviour, in the face of the seemingly overriding passions for self-preservation and self-interest. Smith's answer, at considerable length, is the presence within each of us of an "inner man" who plays the role of the "impartial spectator," approving or condemning our own and others' actions with a voice impossible to disregard. (The theory may sound less naive if the question is reformulated to ask how instinctual drives are socialized through the superego.)

Smith's view of human nature

The thesis of the impartial spectator, however, conceals a more important aspect of the book. Smith saw humans as creatures driven by passions and at the same time self-regulated by their ability to reason and—no less important—by their capacity for sympathy. This duality serves both to pit individuals against one another and to provide them with the rational and moral faculties to create institutions by which the internecine struggle can be mitigated and even turned to the common good. He wrote in his *Moral Sentiments* the famous observation that he was to repeat later in *The Wealth of Nations:* that self-seeking men are often "led by an invisible hand . . . without knowing it, without intending it, [to] advance the interest of the society."

It should be noted that scholars have long debated whether *Moral Sentiments* complemented or was in conflict with *The Wealth of Nations,* which followed it. At one level there is a seeming clash between the theme of social morality contained in the first and the largely amoral explication of the economic system in the second. On the other hand, the first book can also be seen as an explanation of the manner in which individuals are socialized to become the market-oriented and class-bound actors that set the economic system into motion.

Travels on the Continent. The *Theory* quickly brought Smith wide esteem and in particular attracted the attention of Charles Townshend, himself something of an amateur economist, a considerable wit, and somewhat less of a statesman, whose fate it was to be the chancellor of the exchequer responsible for the measures of taxation that ultimately provoked the American Revolution. Townshend had recently married and was searching for a tutor for his stepson and ward, the young Duke of Buccleuch. Influenced by the strong recommendations of Hume and his own admiration for *The Theory of Moral Sentiments,* he approached Smith to take the charge.

The terms of employment were lucrative (an annual salary of £300 plus traveling expenses and a pension of £300 a year thereafter), considerably more than Smith had earned as a professor. Accordingly, Smith resigned his Glasgow post in 1763 and set off for France the next year as the tutor of the young duke. They stayed mainly in Toulouse, where Smith began working on a book (eventually to be *The Wealth of Nations*) as an antidote to the excruciating boredom of the provinces. After 18 months of ennui he was rewarded with a two-month sojourn in Geneva, where he met Voltaire, for whom he had the profoundest respect, thence to Paris, where Hume, then secretary to the British embassy, introduced Smith to the great literary salons of the French Enlightenment. There he met a group of social reformers and theorists headed by François Quesnay, who called themselves *les économistes* but are known in history as the physiocrats. There is some controversy as to the precise degree of influence the physiocrats exerted on Smith, but it is known that he thought sufficiently well of Quesnay to have considered dedicating *The Wealth of Nations* to him, had not the French economist died before publication.

The stay in Paris was cut short by a shocking event. The younger brother of the Duke of Buccleuch, who had joined them in Toulouse, took ill and perished despite Smith's frantic ministrations. Smith and his charge immediately returned to London. Smith worked in London until the spring of 1767 with Lord Townshend, a period during which he was elected a fellow of the Royal Society and broadened still further his intellectual circle to include Edmund Burke, Samuel Johnson, Edward Gibbon, and perhaps Benjamin Franklin. Late that year he returned to Kirkcaldy, where the next six years were spent dictating and reworking *The Wealth of Nations,* followed by another stay of three years in London, where the work was finally completed and published in 1776.

The Wealth of Nations. Despite its renown as the first great work in political economy, *The Wealth of Nations* is in fact a continuation of the philosophical theme begun in *The Theory of Moral Sentiments.* The ultimate problem to which Smith addresses himself is how the inner struggle between the passions and the "impartial spectator"—explicated in *Moral Sentiments* in terms of the single individual—works its effects in the larger arena of history itself, both in the long-run evolution of society and in terms of the immediate characteristics of the stage of history typical of Smith's own day.

Stages of social organization

The answer to this problem enters in Book V, in which Smith outlines the four main stages of organization through which society is impelled, unless blocked by deficiencies of resources, wars, or bad policies of government: the original "rude" state of hunters, a second stage of nomadic agriculture, a third stage of feudal or manorial "farming," and a fourth and final stage of commercial interdependence.

It should be noted that each of these stages is accompanied by institutions suited to its needs. For example, in the age of the huntsman, "there is scarce any property . . . ; so there is seldom any established magistrate or any regular administration of justice." With the advent of flocks there emerges a more complex form of social organization, comprising not only "formidable" armies but the central institution of private property with its indispensable buttress of law and order as well. It is the very essence of Smith's thought that he recognized this institution, whose social usefulness he never doubted, as an instrument for the protection of privilege, rather than one to be justified in terms of natural law: "Civil government," he wrote, "so far as it is instituted for the security of property, is in reality instituted for the defence of the rich against the poor, or of those who have some property against those who have none at all." Finally, Smith describes the evolution through feudalism into a stage of society requiring new institutions, such as market-determined rather than guild-determined wages and free rather than government-constrained enterprise. This later became known as laissez-faire capitalism; Smith called it the system of perfect liberty.

There is an obvious resemblance between this succession of changes in the material basis of production, each bringing its requisite alterations in the superstructure of laws and civil institutions, and the Marxian conception of history. Though the resemblance is indeed remarkable, there is also a crucial difference: in the Marxian scheme the engine of evolution is ultimately the struggle between contending classes, whereas in Smith's philosophical history the primal moving agency is "human nature" driven by the desire for self-betterment and guided (or misguided) by the faculties of reason.

Society and the "invisible hand." The theory of historical evolution, although it is perhaps the binding conception of *The Wealth of Nations,* is subordinated within the work itself to a detailed description of how the "invisible hand" actually operates within the commercial, or final, stage of society. This becomes the focus of Books I and II, in which Smith undertakes to elucidate two questions. The first is how a system of perfect liberty, operating under the drives and constraints of human nature and intelligently designed institutions, will give rise to an orderly society. The question, which had already been considerably elucidated by earlier writers, required both an explanation of the underlying orderliness in the pricing of individual commodities and an explanation of the "laws" that regulated the division of the entire "wealth" of the nation (which Smith saw as its annual production of goods and services) among the three great claimant classes—labourers, landlords, and manufacturers.

This orderliness, as would be expected, was produced by the interaction of the two aspects of human nature, its response to its passions and its susceptibility to reason and sympathy. But whereas *The Theory of Moral Sentiments* had relied mainly on the presence of the "inner man" to provide the necessary restraints to private action, in *The Wealth of Nations* one finds an institutional mechanism that acts to reconcile the disruptive possibilities inherent in a blind obedience to the passions alone. This protective mechanism is competition, an arrangement by which the passionate desire for bettering one's condition—"a desire that comes with us from the womb, and never leaves us until we go into the grave"—is turned into a socially beneficial agency by pitting one person's drive for self-betterment against another's.

The role of competition

It is in the unintended outcome of this competitive struggle for self-betterment that the invisible hand regulating the economy shows itself, for Smith explains how mutual vying forces the prices of commodities down to their "natural" levels, which correspond to their costs of production. Moreover, by inducing labour and capital to move from less to more profitable occupations or areas, the competitive mechanism constantly restores prices to these "natural" levels despite short-run aberrations. Finally, by explaining that wages and rents and profits (the constituent parts of the costs of production) are themselves subject to this same discipline of self-interest and competition, Smith not only provided an ultimate rationale for these "natural" prices but also revealed an underlying orderliness in the distribution of income itself among workers, whose recompense was their wages; landlords, whose income was their rents; and manufacturers, whose reward was their profits.

Economic growth. Smith's analysis of the market as a self-correcting mechanism was impressive. But his purpose was more ambitious than to demonstrate the self-adjusting properties of the system. Rather, it was to show that, under the impetus of the acquisitive drive, the annual flow of national wealth could be seen steadily to grow.

Division of labour

Smith's explanation of economic growth, although not neatly assembled in one part of *The Wealth of Nations,* is quite clear. The core of it lies in his emphasis on the division of labour (itself an outgrowth of the "natural" propensity to trade) as the source of society's capacity to increase its productivity. *The Wealth of Nations* opens with a famous passage describing a pin factory in which 10 persons, by specializing in various tasks, turn out 48,000 pins a day, compared with the few, perhaps only 1, that each could have produced alone. But this all-important division of labour does not take place unaided. It can occur only after the prior accumulation of capital (or stock, as Smith calls it), which is used to pay the additional workers and to buy tools and machines.

The drive for accumulation, however, brings problems. The manufacturer who accumulates stock needs more labourers (since labour-saving technology has no place in Smith's scheme), and in attempting to hire them he bids up their wages above their "natural" price. Consequently his profits begin to fall, and the process of accumulation is in danger of ceasing. But now there enters an ingenious mechanism for continuing the advance. In bidding up the price of labour, the manufacturer inadvertently sets into motion a process that increases the supply of labour, for "the demand for men, like that for any other commodity, necessarily regulates the production of men." Specifically, Smith had in mind the effect of higher wages in lessening child mortality. Under the influence of a larger labour supply, the wage rise is moderated and profits are maintained; the new supply of labourers offers a continuing opportunity for the manufacturer to introduce a further division of labour and thereby add to the system's growth.

Here then was a "machine" for growth—a machine that operated with all the reliability of the Newtonian system with which Smith was quite familiar. Unlike the Newtonian system, however, Smith's growth machine did not depend for its operation on the laws of nature alone. Human nature drove it, and human nature was a complex rather than a simple force. Thus, the wealth of nations would grow only if individuals, through their governments, did not inhibit this growth by catering to the pleas for special privilege that would prevent the competitive system from exerting its benign effect. Consequently, much of *The Wealth of Nations,* especially Book IV, is a polemic against the restrictive measures of the "mercantile system" that favoured monopolies at home and abroad. Smith's system of "natural liberty," he is careful to point out, accords with the best interests of all but will not be put into practice if government is entrusted to, or heeds, "the mean rapacity, the monopolizing spirit of merchants and manufacturers, who neither are, nor ought to be, the rulers of mankind."

The attack on mercantilism

The Wealth of Nations is therefore far from the ideological tract it is often supposed to be. Although Smith preached laissez-faire (with important exceptions), his argument was directed as much against monopoly as government; and although he extolled the social results of the acquisitive process, he almost invariably treated the manners and maneuvers of businessmen with contempt. Nor did he see the commercial system itself as wholly admirable. He wrote with discernment about the intellectual degradation of the worker in a society in which the division of labour has proceeded very far; for by comparison with the alert intelligence of the husbandman, the specialized worker "generally becomes as stupid and ignorant as it is possible for a human being to become."

In all of this, it is notable that Smith was writing in an age of preindustrial capitalism. He seems to have had no real presentiment of the gathering Industrial Revolution, harbingers of which were visible in the great ironworks only a few miles from Edinburgh. He had nothing to say about large-scale industrial enterprise, and the few remarks in *The Wealth of Nations* concerning the future of joint-stock companies (corporations) are disparaging. Finally, one should bear in mind that, if growth is the great theme of *The Wealth of Nations,* it is not unending growth. Here and there in the treatise are glimpses of a secularly declining rate of profit; and Smith mentions as well the prospect that when the system eventually accumulates its "full complement of riches"—all the pin factories, so to speak, whose output could be absorbed—economic decline would begin, ending in an impoverished stagnation.

The Wealth of Nations was received with admiration by Smith's wide circle of friends and admirers, although it was by no means an immediate popular success. The work finished, Smith went into semiretirement. The year following its publication he was appointed commissioner both of customs and of salt duties for Scotland, posts that brought him £600 a year. He thereupon informed his former charge that he no longer required his pension, to which Buccleuch replied that his sense of honour would never allow him to stop paying it. Smith was therefore quite well off in the final years of his life, which were spent mainly in Edinburgh with occasional trips to London or Glasgow (which appointed him a rector of the university). The years passed quietly, with several revisions of both major books but with no further publications. On July 17, 1790, at the age of 67, full of honours and recognition, Smith died; he was buried in the churchyard at Canongate with a simple monument stating that Adam Smith, author of *The Wealth of Nations,* was buried there.

Later life

Beyond the few facts of his life, which can be embroidered only in detail, exasperatingly little is known about the man. Smith never married, and almost nothing is known of his personal side. Moreover, it was the custom of his time to destroy rather than to preserve the private files of illustrious men, with the unhappy result that much of Smith's unfinished work, as well as his personal papers, was destroyed (some as late as 1942). Only one portrait of Smith survives, a profile medallion by Tassie; it gives a glimpse of the older man with his somewhat heavy-lidded eyes, aquiline nose, and a hint of a protrusive lower lip. "I am a beau in nothing but my books," Smith once told a friend to whom he was showing his library of some 3,000 volumes.

Personal qualities

From various accounts, he was also a man of many peculiarities, which included a stumbling manner of speech (until he had warmed to his subject), a gait described as "vermicular," and above all an extraordinary and even comic absence of mind. On the other hand, contemporaries wrote of a smile of "inexpressible benignity," and of his political tact and dispatch in managing the sometimes acerbic business of the Glasgow faculty.

Certainly he enjoyed a high measure of contemporary fame; even in his early days at Glasgow his reputation attracted students from nations as distant as Russia, and his later years were crowned not only with expressions of admiration from many European thinkers but by a growing recognition among British governing circles that his work provided a rationale of inestimable importance for practical economic policy.

Over the years, Smith's lustre as a social philosopher has escaped much of the weathering that has affected the reputations of other first-rate political economists. Although he was writing for his generation, the breadth of his knowl-

edge, the cutting edge of his generalizations, the boldness of his vision, have never ceased to attract the admiration of all social scientists, and in particular economists. Couched in the spacious, cadenced prose of his period, rich in imagery and crowded with life, *The Wealth of Nations* projects a sanguine but never sentimental image of society. Never so finely analytic as David Ricardo nor so stern and profound as Karl Marx, Smith is the very epitome of the Enlightenment: hopeful but realistic, speculative but practical, always respectful of the classical past but ultimately dedicated to the great discovery of his age—progress.

BIBLIOGRAPHY. The complete works have appeared in a definitive edition, "The Glasgow Edition of the Works and Correspondence of Adam Smith," 6 vol. in 7 (1976–83), including vol. 1, *The Theory of Moral Sentiments*, ed. by D.D. RAPHAEL and A.L. MACFIE (1976), vol. 2, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vol., ed. by R.H. CAMPBELL and A.S. SKINNER, vol. 3, *Essays on Philosophical Subjects*, ed. by W.P.D. WIGHTMAN and J.C. BRYCE (1980), which contains the interesting "The History of Astronomy," vol. 4, *Lectures on Rhetoric and Belles Lettres*, ed. by J.C. BRYCE (1983), and vol. 5, *Lectures on Jurisprudence*, ed. by R.L. MEEK, D.D. RAPHAEL, and P.G. STEIN (1978). For the nonspecialist, ROBERT L. HEILBRONER (ed.), *The Essential Adam Smith* (1986), offers fairly extensive readings and short discussions of Smith's main works.

Among biographical works are JOHN RAE, *Life of Adam Smith* (1895, reprinted 1965); WILLIAM R. SCOTT, *Adam Smith as Student and Professor* (1937, reprinted 1965), including "An Early Draft of Part of *The Wealth of Nations*," various documents, and correspondence; and DUGALD STEWART, *Biographical Memoirs of Adam Smith...*, vol. 10 in *The Collected Works of Dugald Stewart* (1858, reprinted 1966).

DONALD WINCH, *Adam Smith's Politics: An Essay in Historiographic Revision* (1978), reinterprets Smith's place in the history of economic and political thought. ANDREW S. SKINNER and THOMAS WILSON, *Essays on Adam Smith* (1975), contains discussion by well-known scholars of various aspects of Smith's work. KNUD HAAKONSSEN, *The Science of a Legislator: The Natural Jurisprudence of David Hume and Adam Smith* (1981), compares their philosophical systems. Useful articles include ADOLPH LOWE, "The Classical Theory of Economic Growth," *Social Research*, 21(2):127–158 (Summer 1954); NATHAN ROSENBERG, "Adam Smith on the Division of Labour: Two Views or One?" *Economica*, 32(126):127–139 (May 1965), and "Some Institutional Aspects of the *Wealth of Nations*," *The Journal of Political Economy*, 68(6):557–570 (December 1960); JOSEPH J. SPENGLER, "Adam Smith's Theory of Economic Growth," *Southern Economic Journal*, 25(4):397–415, 26(1):1–12 (April and July 1959); the entry by JACOB VINER, "Adam Smith," in DAVID L. SILLS (ed.), *International Encyclopaedia of the Social Sciences*, vol. 14, pp. 322–329 (1968); and the entry by ANDREW S. SKINNER, "Adam Smith," in *The New Palgrave: A Dictionary of Economics*, ed. by JOHN EATWELL MURRAY MILGATE and PETER NEWMAN, vol. 4 (1987), pp. 357–375, with a bibliography.

(R.L.He.)

The Social Sciences

The social sciences, which deal with human behaviour in its social and cultural aspects, include the following disciplines: anthropology, sociology, economics, political science, and the study of international relations. Also frequently included are social and economic geography and those areas of education that deal with the social contexts of learning and the relation of the school to the social order. The study of comparative law may also be regarded as a part of the social sciences, although it is ordinarily pursued in schools of law rather than in departments or schools containing most of the other social sciences.

Since the 1950s the term behavioral sciences has often been applied to the disciplines designated as the social sciences. Those who favour this term do so in part because these disciplines are thus brought closer to some of the sciences, such as physical anthropology and physiological psychology, which also deal with human behaviour. Whether the term *behavioral sciences* will in time supplant *social sciences* or whether it will fade away is impossible to say. For the purposes of this article, the two terms may be considered synonymous.

This article is concerned with the social sciences as vital elements in the aftermath of the two great revolutions, the political and industrial, which opened the 19th century, with the pattern the social sciences assumed in that century, and with their development in the 20th century.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/31, 10/36, and 10/41, and the *Index.*

The article is divided into the following sections:

History of the social sciences 317
- Heritage of the Middle Ages and the Renaissance 317
- Heritage of the Enlightenment 317
- The 19th century 318
  - Major themes resulting from democratic and industrial change
  - New ideologies
  - New intellectual and philosophical tendencies
  - Development of the separate disciplines
- The 20th century 322
  - Marxist influences
  - Freudian influences
  - Specialization and cross-disciplinary approaches
  - Nature of the research
  - Theoretical modes

Anthropology 326
- Overview 326
- History of anthropology 326
  - Fieldwork
  - Social and cultural anthropology
  - American anthropology since the 1950s
  - European anthropology since the 1950s
- Cultural anthropology 328
  - American cultural anthropology
  - French theoretical contributions
  - The configurational approach
  - Cultural change and adaptation
  - Culture and the humanities
- Linguistic anthropology 330
- Psychological anthropology 331

Sociology 331
- Historical development of sociology 331
  - Founding the discipline
  - Replacing Darwinist determinism
- Early schools of thought 332
  - Early functionalism
  - The functionalist-conflict debate
  - Rising segmentation of the discipline
  - Major modern developments
- Methodological considerations in sociology 334
  - Methodological development in contemporary sociology
  - Ecological patterning
  - Experiments
  - Statistics and mathematical analysis
  - Data collection
  - National methodological preferences
- Status of contemporary sociology 336
  - Academic status
  - Scientific status
  - Current trends
  - Emerging roles for sociologists
- Related fields 338
  - Social psychology
  - Criminology

Economics 343
- Historical development of economics 343
  - Analysis of the market
  - Construction of a system
  - Marxism
  - The marginalists
  - The critics
  - Keynesian economics
  - Postwar developments
- Methodological considerations in contemporary economics 346
  - Methods of inference
  - Testing theories
  - Microeconomics
  - Macroeconomics
- Fields of contemporary economics 349
  - Public finance
  - Money
  - International economics
  - Labour
  - Industrial organization
  - Agriculture
  - Growth and development
  - Mathematical economics
  - Econometrics

Political science 352
- Historical development of political science 352
  - Early trends
  - Juristic influences
  - Developments in the United States
  - Developments in Europe
- Methodological considerations in contemporary political science 354
  - Behavioralism
  - Systems analysis
  - Interest groups, elites, and political parties
  - Analysis of political attitudes and voting behaviour
- Current trends 355

Study of international relations 356
- Historical development of the study of international relations 356
  - Between the two world wars
  - The postwar ascendancy of political realism
  - The behavioral decade: mid-1950s to mid-1960s
  - Contemporary perspectives in international relations
- Relations between scholarship and action in foreign affairs 360

Comparative law 360
- Historical development of comparative law 360
  - Ancient roots of law
  - Role of judges
  - 19th-century beginnings
  - International efforts
- Methodological considerations in contemporary comparative law 361
  - Microcomparison
  - Macrocomparison
  - Classification of families of law
- Purposes of comparative law 362
  - Historical and cultural comparisons
  - Commercial uses
  - Aid to national law
  - Use in international law

Bibliography 363

History of the social sciences

Although, strictly speaking, the social sciences do not precede the 19th century—that is, as distinct and recognized disciplines of thought—one must go back farther in time for the origins of some of their fundamental ideas and objectives. In the largest sense, the origins go all the way back to the ancient Greeks and their rationalist inquiries into the nature of man, state, and morality. The heritage of both Greece and Rome is a powerful one in the history of social thought as it is in so many other areas of Western society. Very probably, apart from the initial Greek determination to study all things in the spirit of dispassionate and rational inquiry, there would be no social sciences today. True, there have been long periods of time, as during the Western Middle Ages, when the Greek rationalist temper was lacking. But the recovery of this temper, through texts of the great classical philosophers, is the very essence of the Renaissance and the Age of Reason in modern European history. With the Age of Reason, in the 17th and 18th centuries, one may begin.

HERITAGE OF THE MIDDLE AGES AND THE RENAISSANCE

The same impulses that led men in that age to explore the earth, the stellar regions, and the nature of matter led them also to explore the institutions around them: state, economy, religion, morality; above all, the nature of man himself. It was the fragmentation of medieval philosophy and theory, and, with this, the shattering of the medieval world view that had lain deep in thought until about the 16th century, that was the immediate basis of the rise of the several strands of specialized thought that were to become in time the social sciences.

Adverse effects of medieval theology

Medieval theology, especially as it appears in St. Thomas Aquinas' *Summa theologiae,* contained and fashioned syntheses from ideas about man and society—ideas indeed that may be seen to be political, social, economic, anthropological, and geographical in their substance. But it was partly this close relation between medieval theology and ideas of the social sciences that accounts for the longer time it took these ideas—by comparison with the ideas of the physical sciences—to achieve what one would today call scientific character. From the time of the great Roger Bacon in the 13th century, there were at least some rudiments of physical science that were largely independent of medieval theology and philosophy. Historians of physical science have no difficulty in tracing the continuation of this experimental tradition, primitive and irregular though it was by later standards, throughout the Middle Ages. Side by side with the kinds of experiment made notable by Roger Bacon were impressive changes in technology through the medieval period and then, in striking degree, in the Renaissance. Efforts to improve agricultural productivity; the rising utilization of gunpowder, with consequent development of guns and the problems that they presented in ballistics; growing trade, leading to increased use of ships and improvements in the arts of navigation, including use of telescopes; and the whole range of such mechanical arts in the Middle Ages and Renaissance as architecture, engineering, optics, and the construction of watches and clocks—all of this put a high premium on a pragmatic and operational understanding of at least the simpler principles of mechanics, physics, astronomy, and, in time, chemistry.

In short, by the time of Copernicus and Galileo in the 16th century, a fairly broad substratum of physical science existed, largely empirical but not without theoretical implications on which the edifice of modern physical science could be built. It is notable that the empirical foundations of physiology were being established in the studies of the human body being conducted in medieval schools of medicine and, as the career of Leonardo da Vinci so resplendently illustrates, among artists of the Renaissance, whose interest in accuracy and detail of painting and sculpture led to their careful studies of human anatomy.

Very different was the beginning of the social sciences. In the first place, the church, throughout the Middle Ages and even into the Renaissance and Reformation, was much more attentive to what scholars wrote and thought about man's mind and his behaviour in society than it was toward what was being studied and written in the physical sciences. From the church's point of view, while it might be important to see to it that thought on the physical world corresponded as far as possible to what Scripture said—witnessed, for example, in the famous questioning of Galileo—it was far more important that such correspondence exist in matters affecting the nature of man, his mind, spirit, and soul. Nearly all the subjects and questions that would form the bases of the social sciences in later centuries were tightly woven into the fabric of medieval scholasticism, and it was not easy for even the boldest minds to break this fabric.

Then, when the hold of scholasticism did begin to wane, two fresh influences, equally powerful, came on the scene to prevent anything comparable to the pragmatic and empirical foundations of the physical sciences from forming in the study of man and society. The first was the immense appeal of the Greek classics during the Renaissance, especially those of the philosophers Plato and Aristotle. A great deal of social thought during the Renaissance was little more than gloss or commentary on the Greek classics. One sees this throughout the 15th and 16th centuries.

Adverse effects of reverence for classics and Cartesianism

Second, in the 17th century appeared the powerful influence of the philosopher René Descartes. Cartesianism, as his philosophy was called, declared that the proper approach to understanding of the world, including man and society, was through a few simple, fundamental ideas of reality and, then, rigorous, almost geometrical deduction of more complex ideas and eventually of large, encompassing theories, from these simple ideas, all of which, Descartes insisted, were the stock of common sense—the mind that is common to all human beings at birth. It would be hard to exaggerate the impact of Cartesianism on social and political and moral thought during the century and a half following publication of his *Discourse on Method* and his *Meditations.* Through the Age of Reason and down through the Enlightenment in the later 18th century, the spell of Cartesianism was cast on nearly all those who were concerned with the problems of the nature of man and society.

Both of these great influences, reverence for the classics and fascination with the geometrical-deductive procedures advocated by Descartes must be seen from today's vantage point as among the major influences retarding the development of a science of society comparable to the science of the physical world. It is not as though data were not available in the 17th and 18th centuries. The emergence of the national state carried with it evergrowing bureaucracies concerned with gathering information, chiefly for taxation, census, and trade purposes, which might have been employed in much the same way that physical scientists employed their data. The voluminous and widely published accounts of the great voyages that had begun in the 15th century, the records of soldiers, explorers, and missionaries who perforce had been brought into often long and close contact with primitive and other non-Western peoples, provided still another great reservoir of data, all of which might have been utilized in scientific ways as such data were to be utilized a century or two later in the social sciences. Such, however, was the continuing spell cast by the texts of the classics and by the strictly rationalistic, overwhelmingly deductive procedures of the Cartesians that, down until the beginning of the 19th century, these and other empirical materials were used, if at all, solely for illustrative purposes in the writings of the social philosophers.

HERITAGE OF THE ENLIGHTENMENT

There is also the fact that, especially in the 18th century, reform and even revolution were often in the air. The purpose of a great many social philosophers was by no means restricted to philosophic, much less scientific, understanding of man and society. The dead hand of the Middle Ages seemed to many vigorous minds in western Europe the principal force to be combatted, through critical reason, enlightenment, and, where necessary, major reform or revolution. One may properly account a great deal of this new spirit to the rise of humanitarianism in

modern Europe and in other parts of the world and to the spread of literacy, the rise in the standard of living, and the recognition that poverty and oppression need not be the fate of the masses. The fact remains, however, that social reform and social science have different organizing principles, and the very fact that for a long time, down indeed through a good part of the 19th century, social reform and social science were regarded as pretty much the same thing could not have helped but retard the development of the latter.

Nevertheless, it would be wrong to discount the significant contributions to the social sciences that were made during the 17th and 18th centuries. The first and greatest of these was the spreading ideal of a science of society, an ideal fully as widespread by the 18th century as the ideal of a physical science. Second was the rising awareness of the multiplicity and variety of human experience in the world. Ethnocentrism and parochialism, as states of mind, were more and more difficult for educated people to maintain given the immense amount of information about—or, more important, interest in—non-Western peoples, the results of trade and exploration. Third was the spreading sense of the social or cultural character of human behaviour in society—that is, its purely historical or conventional, rather than biological, basis. A science of society, in short, was no mere appendage of biology but was instead a distinct discipline, or set of disciplines, with its own distinctive subject matter.

Ideas of structure and developmental change

To these may be added two other very important contributions of the 17th and 18th centuries, each of great theoretical importance. The first was the idea of structure. First seen in the writings of such philosophers as Hobbes, Locke, and Rousseau with reference to the political structure of the state, it had spread by the mid-18th century to highlight the economic writings of the Physiocrats and Adam Smith. The idea of structure can also be seen in certain works relating to man's psychology and, at opposite reach, to the whole of civil society. The ideas of structure that were borrowed from both the physical and biological sciences were fundamental to the conceptions of political, economic, and social structure that took shape in the 17th and 18th centuries. And these conceptions of structure have in many instances, subject only to minor changes, come down to 20th-century social science.

The second major theoretical idea was that of developmental change. Its ultimate roots in Western thought, like those indeed of the whole idea of structure, go back to the Greeks, if not earlier. But it is in the 18th century, above all others, that the philosophy of developmentalism took shape, forming a preview, so to speak, of the social evolutionism of the next century. What was said by such writers as Condorcet, Rousseau, and Adam Smith was that the present is an outgrowth of the past, the result of a long line of development in time, and, furthermore, a line of development that has been caused, not by God or fortuitous factors, but by conditions and causes immanent in human society. Despite a fairly widespread belief that the idea of social development is a product of prior discovery of biological evolution, the facts are the reverse. Well before any clear idea of genetic speciation existed in European biology, there was a very clear idea of what might be called social speciation—that is, the emergence of one institution from another in time and of the whole differentiation of function and structure that goes with this emergence.

As has been suggested, these and other seminal ideas were contained for the most part in writings, the primary function of which was attack on the existing order of government and society in western Europe. Another way of putting the matter is to say that they were clear and acknowledged parts of political and social idealism—using that word in its largest sense. Hobbes, Locke, Rousseau, Montesquieu, Adam Smith, and other major philosophers had as vivid and energizing sense of the ideal—ideal state, ideal economy, ideal civil society—as any earlier utopian writer. These men were, without exception, committed to visions of the good or ideal society. Their interest in the "natural"—that is, natural morality, religion, economy, or education, in contrast to the merely conventional and historically derived—sprang as much from the desire to hold a glass up to a surrounding society that they disliked as from any dispassionate urge simply to find out what man and society are made of. The fact remains, however, that the ideas that were to prove decisive in the 19th century, so far as the social sciences were concerned, arose during the two centuries preceding.

THE 19TH CENTURY

The fundamental ideas, themes, and problems of the social sciences in the 19th century are best understood as responses to the problem of order that was created in men's minds by the weakening of the old order, or European society, under the twin blows of the French Revolution and the Industrial Revolution. The breakup of the old order—an order that had rested on kinship, land, social class, religion, local community, and monarchy—set free, as it were, the complex elements of status, authority, and wealth that had been for so long consolidated. In the same way that the history of 19th-century politics, industry, and trade is basically about the practical efforts of human beings to reconsolidate these elements, so the history of 19th-century social thought is about theoretical efforts to reconsolidate them—that is, to give them new contexts of meaning.

Effects of the democratic and industrial revolutions

In terms of the immediacy and sheer massiveness of impact on human thought and values, it would be difficult to find revolutions of comparable magnitude in human history. The political, social, and cultural changes that began in France and England at the very end of the 18th century spread almost immediately through Europe and the Americas in the 19th century and then on to Asia, Africa, and Oceania in the 20th. The effects of the two revolutions, the one overwhelmingly democratic in thrust, the other industrial-capitalist, have been to undermine, shake, or topple institutions that had endured for centuries, even millennia, and with them systems of authority, status, belief, and community.

It is easy today to deprecate the suddenness, the cataclysmic nature, the overall revolutionary effect of these two changes and to seek to subordinate results to longer, deeper tendencies of more gradual change in western Europe. But as many recent historians have pointed out, there was to be seen, and seen by a great many sensitive minds of that day, a dramatic and convulsive quality to the changes that cannot properly be subsumed to the slower processes of continuous evolutionary change. What is crucial, in any event, from the point of view of the history of the social thought of the period, is how the changes were actually envisaged at the time. By a large number of social philosophers and social scientists, in all spheres, those changes were regarded as nothing less than of earthquake intensity.

The coining or redefining of words is an excellent indication of men's perceptions of change in a given historical period. A large number of words taken for granted today came into being in the period marked by the final decade or two of the 18th century and the first quarter of the 19th. Among these are: industry, industrialist, democracy, class, middle class, ideology, intellectual, rationalism, humanitarian, atomistic, masses, commercialism, proletariat, collectivism, equalitarian, liberal, conservative, scientist, utilitarian, bureaucracy, capitalism, and crisis. Some of these words were invented; others reflect new and very different meanings given to old ones. All alike bear witness to the transformed character of the European social landscape as this landscape loomed up to the leading minds of the age. And all these words bear witness too to the emergence of new social philosophies and, most pertinent to the subject of this article, the social sciences as they are known today.

Major themes resulting from democratic and industrial change. It is illuminating to mention a few of the major themes in social thought in the 19th century that were almost the direct results of the democratic and industrial revolutions. It should be borne in mind that these themes are to be seen in the philosophical and literary writing of the age as well as in social thought.

First, there was the great increase in population. Between

Effects of population growth

1750 and 1850 the population of Europe went from 140,000,000 to 266,000,000; in the world from 728,000,000 to well over 1,000,000,000. It was an English clergyman-economist, Thomas Malthus, who, in his famous *Essay on Population,* first marked the enormous significance to human welfare of this increase. With the diminution of historic checks on population growth, chiefly those of high mortality rates—a diminution that was, as Malthus realized, one of the rewards of technical progress—there were no easily foreseeable limits to growth of population. And such growth, he stressed, could only upset the traditional balance between population, which Malthus described as growing at geometrical rate, and food supply, which he declared could grow only at arithmetical rate. Not all social scientists in the century took the pessimistic view of the matter that Malthus did but few if any were indifferent to the impact of explosive increase in population on economy, government, and society.

Second, there was the condition of labour. It may be possible to see this condition in the early 19th century as in fact better than the condition of the rural masses at earlier times. But the important point is that to a large number of writers in the 19th century it seemed worse and was defined as worse. The wrenching of large numbers of people from the older and protective contexts of village, guild, parish, and family, and their massing in the new centres of industry, forming slums, living in common squalor and wretchedness, their wages generally behind cost of living, their families growing larger, their standard of living becoming lower, as it seemed—all of this is a frequent theme in the social thought of the century. Economics indeed became known as the "dismal science," because economists, from David Ricardo to Karl Marx, could see little likelihood of the condition of labour improving under capitalism.

Third, there was the transformation of property. Not only was more and more property to be seen as industrial—manifest in the factories, business houses, and workshops of the period—but also the very nature of property was changing. Whereas for most of the history of mankind property had been "hard," visible only in concrete possessions—land and money—now the more intangible kinds of property such as shares of stock, negotiable equities of all kinds, and bonds were assuming ever greater influence in the economy. This led, as was early realized, to the dominance of financial interests, to speculation, and to a general widening of the gulf between the propertied and the masses. The change in the character of property made easier the concentration of property, the accumulation of immense wealth in the hands of a relative few, and, not least, the possibility of economic domination of politics and culture. It should not be thought that only socialists saw property in this light. From Edmund Burke through Auguste Comte, Frédéric Le Play, and John Stuart Mill down to Karl Marx, Max Weber, and Émile Durkheim, one finds conservatives and liberals looking at the impact of this change in analogous ways.

Effects of urbanization and technological change

Fourth, there was urbanization—the sudden increase in the number of towns and cities in western Europe and the increase in number of persons living in the historic towns and cities. Whereas in earlier centuries, the city had been regarded almost uniformly as a setting of civilization, culture, and freedom of mind, now one found more and more writers aware of the other side of cities: the atomization of human relationships, broken families, the sense of the mass, of anonymity, alienation, and disrupted values. Sociology particularly among the social sciences turned its attention to the problems of urbanization. The contrast between the more organic type of community found in rural areas and the more mechanical and individualistic society of the cities is a basic contrast in sociology, one that was given much attention by such pioneers in Europe as the French sociologists Frédéric Le Play and Émile Durkheim; the German sociologists Ferdinand Tönnies, Georg Simmel, and Max Weber; the Belgian statistician Adolphe Quetelet; and, in America, by the sociologists Charles H. Cooley and Robert E. Park.

Fifth, there was technology. With the spread of mechanization, first in the factories, then in agriculture, social thinkers could see possibilities of a rupture of the historic relation between man and nature, between man and man, even between man and God. To thinkers as politically different as Thomas Carlyle and Karl Marx, technology seemed to lead to dehumanization of the worker and to exercise of a new kind of tyranny over human life. Marx, though, far from despising technology, thought the advent of socialism would counteract all this. Alexis de Tocqueville declared that technology, and especially technical specialization of work, was more degrading to man's mind and spirit than even political tyranny. It was thus in the 19th century that the opposition to technology on moral, psychological, and aesthetic grounds first made its appearance in Western thought.

Sixth, there was the factory system. The importance of this to 19th-century thought has been intimated above. Suffice it to add that along with urbanization and spreading mechanization, the system of work whereby masses of workers left home and family to work long hours in the factories became a major theme of social thought as well as of social reform.

Seventh, and finally, mention is to be made of the development of political masses—that is, the slow but inexorable widening of franchise and electorate through which ever larger numbers of persons became aware of themselves as voters and participants in the political process. This too is a major theme in social thought, to be seen most luminously perhaps in Tocqueville's *Democracy in America,* a classic written in the 1830s that took not merely America but democracy everywhere as its subject. Tocqueville saw the rise of the political masses, more especially the immense power that could be wielded by the masses, as the single greatest threat to individual freedom and cultural diversity in the ages ahead.

Effects of the rise of the masses

These, then, are the principal themes in the 19th-century writing that may be seen as direct results of the two great revolutions. As themes, they are to be found not only in the social sciences but, as noted above, in a great deal of the philosophical and literary writing of the century. In their respective ways, the philosophers Hegel, Coleridge, and Emerson were as struck by the consequences of the revolutions as were any social scientists. So too were such novelists as Balzac and Dickens.

New ideologies. One other point must be emphasized about these themes. They became, almost immediately in the 19th century, the bases of new ideologies. How men reacted to the currents of democracy and industrialism stamped them conservative, liberal, or radical. On the whole, with rarest exceptions, liberals welcomed the two revolutions, seeing in their forces opportunity for freedom and welfare never before known to mankind. The liberal view of society was overwhelmingly democratic, capitalist, industrial, and, of course, individualistic. The case is somewhat different with conservatism and radicalism in the century. Conservatives, beginning with Edmund Burke, continuing through Hegel and Matthew Arnold down to such minds as John Ruskin later in the century, disliked both democracy and industrialism, preferring the kind of tradition, authority, and civility that had been, in their minds, displaced by the two revolutions. Theirs was a retrospective view, but it was a nonetheless influential one, affecting a number of the central social scientists of the century, among them Auguste Comte and Tocqueville and later Max Weber and Émile Durkheim. The radicals accepted democracy but only in terms of its extension to all areas of society and its eventual annihilation of any form of authority that did not spring directly from the people as a whole. And although the radicals, for the most part, accepted the phenomenon of industrialism, especially technology, they were uniformly antagonistic to capitalism.

These ideological consequences of the two revolutions proved extremely important to the social sciences, for it would be difficult to identify a social scientist in the century—as it would a philosopher or a humanist—who was not, in some degree at least, caught up in ideological currents. In referring to such minds as Saint-Simon, Comte, Le Play among sociologists, to Ricardo, the Frenchman Jean-Baptiste Say, and Marx among economists, to Jeremy

Bentham and John Austin among political scientists, even to anthropologists like the Englishman Edward B. Tylor and the American Lewis Henry Morgan, one has before one men who were engaged not merely in the study of society but also in often strongly partisan ideology. Some were liberals, some conservatives, others radicals. All drew from the currents of ideology that had been generated by the two great revolutions.

New intellectual and philosophical tendencies. It is important also to identify three other powerful tendencies of thought that influenced all of the social sciences. The first is a positivism that was not merely an appeal to science but almost reverence for science; the second, humanitarianism; the third, the philosophy of evolution.

Effects of Positivism

The Positivist appeal of science was to be seen everywhere. The rise of the ideal of science in the Age of Reason was noted above. The 19th century saw the virtual institutionalization of this ideal—possibly even canonization. The great aim was that of dealing with moral values, institutions, and all social phenomena through the same fundamental methods that could be seen so luminously in such areas as physics and biology. Prior to the 19th century, no very clear distinction had been made between philosophy and science, and the term philosophy was even preferred by those working directly with physical materials, seeking laws and principles in the fashion of a Newton or Harvey—that is, by persons whom one would now call scientists.

In the 19th century, in contrast, the distinction between philosophy and science became an overwhelming one. Virtually every area of man's thought and behaviour was thought by a rising number of persons to be amenable to scientific investigation in precisely the same degree that physical data were. More than anyone else, it was Comte who heralded the idea of the scientific treatment of social behaviour. His *Cours de philosophie positive,* published in six volumes between 1830 and 1842, sought to demonstrate irrefutably not merely the possibility but the inevitability of a science of man, one for which Comte coined the word "sociology" and that would do for man the social being exactly what biology had already done for man the biological animal. But Comte was far from alone. There were many in the century to join in his celebration of science for the study of society.

Humanitarianism, though a very distinguishable current of thought in the century, was closely related to the idea of a science of society. For the ultimate purpose of social science was thought by almost everyone to be the welfare of society, the improvement of its moral and social condition. Humanitarianism, strictly defined, is the institutionalization of compassion; it is the extension of welfare and succour from the limited areas in which these had historically been found, chiefly family and village, to society at large. One of the most notable and also distinctive aspects of the 19th century was the constantly rising number of persons, almost wholly from the middle class, who worked directly for the betterment of society. In the many projects and proposals for relief of the destitute, improvement of slums, amelioration of the plight of the insane, the indigent, and imprisoned, and other afflicted minorities could be seen the spirit of humanitarianism at work. All kinds of associations were formed, including temperance associations, groups and societies for the abolition of slavery and of poverty and for the improvement of literacy, among other objectives. Nothing like the 19th-century spirit of humanitarianism had ever been seen before in western Europe—not even in France during the Enlightenment, where interest in mankind's salvation tended to be more intellectual than humanitarian in the strict sense. Humanitarianism and social science were reciprocally related in their purposes. All that helped the cause of the one could be seen as helpful to the other.

Effects of evolutionary theory

The third of the intellectual influences is that of evolution. It affected every one of the social sciences, each of which was as much concerned with the development of things as with their structures. An interest in development was to be found in the 18th century, as noted earlier. But this interest was small and specialized compared with 19th-century theories of social evolution. The impact of Charles Darwin's *Origin of Species,* published in 1859, was of course great and further enhanced the appeal of the evolutionary view of things. But it is very important to recognize that ideas of social evolution had their own origins and contexts. The evolutionary works of such social scientists as Comte, Herbert Spencer, and Marx had been completed, or well begun, before publication of Darwin's work. The important point, in any event, is that the idea or the philosophy of evolution was in the air throughout the century, as profoundly contributory to the establishment of sociology as a systematic discipline in the 1830s as to such fields as geology, astronomy, and biology. Evolution was as permeative an idea as the Trinity had been in medieval Europe.

Development of the separate disciplines. Among the disciplines that formed the social sciences, two contrary, for a time equally powerful, tendencies at first dominated them. The first was the drive toward unification, toward a single, master social science, whatever it might be called. The second tendency was toward specialization of the individual social sciences. If, clearly, it is the second that has triumphed, with the results to be seen in the disparate, sometimes jealous, highly specialized disciplines seen today, the first was not without great importance and must also be examined.

Unification versus specialization

What emerges from the critical rationalism of the 18th century is not, in the first instance, a conception of need for a plurality of social sciences, but rather for a single science of society that would take its place in the hierarchy of the sciences that included the fields of astronomy, physics, chemistry, and biology. When, in the 1820s, Comte wrote calling for a new science, one with man the social animal as the subject, he assuredly had but a single, encompassing science of society in mind—not a congeries of disciplines, each concerned with some single aspect of man's behaviour in society. The same was true of Bentham, Marx, and Spencer. All these minds, and there were many others to join them, saw the study of society as a unified enterprise. They would have scoffed, and on occasion did, at any notion of a separate economics, political science, sociology, and so on. Society is an indivisible thing, they would have argued; so, too, must be the study of society.

It was, however, the opposite tendency of specialization or differentiation that won out. No matter how the century began, or what were the dreams of a Comte, Spencer, or Marx, when the 19th century ended, not one but several distinct, competitive social sciences were to be found. Aiding this process was the development of the colleges and universities. With hindsight it might be said that the cause of universities in the future would have been strengthened, as would the cause of the social sciences, had there come into existence, successfully, a single curriculum, undifferentiated by field, for the study of society. What in fact happened, however, was the opposite. The growing desire for an elective system, for a substantial number of academic specializations, and for differentiation of academic degrees, contributed strongly to the differentiation of the social sciences. This was first and most strongly to be seen in Germany, where, from about 1815 on, all scholarship and science were based in the universities and where competition for status among the several disciplines was keen. But by the end of the century the same phenomenon of specialization was to be found in the United States (where admiration for the German system was very great in academic circles) and, in somewhat less degree, in France and England. Admittedly, the differentiation of the social sciences in the 19th century was but one aspect of a larger process that was to be seen as vividly in the physical sciences and the humanities. No major field escaped the lure of specialization of investigation, and clearly, a great deal of the sheer bulk of learning that passed from the 19th to the 20th century was the direct consequence of this specialization.

Classical economists and socialists

Economics. It was economics that first attained the status of a single and separate science, in ideal at least, among the social sciences. That autonomy and self-regulation that the Physiocrats and Adam Smith had found, or thought they had found, in the processes of wealth, in the operation of prices, rents, interest, and wages during the 18th

century became the basis of a separate and distinctive economics—or, as it was often called, "political economy"—in the 19th. Hence the emphasis upon what came to be widely called laissez-faire. If, as it was argued, the processes of wealth operate naturally in terms of their own built-in mechanisms, then not only should these be studied separately but they should, in any wise polity, be left alone by government and society. This was, in general, the overriding emphasis of such thinkers as David Ricardo, John Stuart Mill, and Nassau William Senior in England, of Frédéric Bastiat and Jean-Baptiste Say in France, and, somewhat later, the Austrian school of Carl Menger. This emphasis is today called "classical" in economics, and it is even now, though with substantial modifications, a strong position in the field.

There were almost from the beginning, however, economists who diverged sharply from this laissez-faire, classical view. In Germany especially there were the so-called historical economists. They proceeded less from the discipline of historiography than from the presuppositions of social evolution, referred to above. Such men as Wilhelm Roscher and Karl Knies in Germany tended to dismiss the assumptions of timelessness and universality regarding economic behaviour that were almost axiomatic among the followers of Adam Smith, and they strongly insisted upon the developmental character of capitalism, evolving in a long series of stages from other types of economy.

Also prominent throughout the century were those who came to be called the Socialists. They too repudiated any notion of timelessness and universality in capitalism and its elements of private property, competition, and profit. Not only was this system but a passing stage of economic developments; it could be—and, as Marx was to emphasize, would be—shortly supplanted by a more humane and also realistic economic system based upon cooperation, the people's ownership of the means of production, and planning that would eradicate the vices of competition and conflict.

Political science. Rivalling economics as a discipline during the century was political science. The line of systematic interest in the state that had begun in modern Europe with Machiavelli, Hobbes, Locke, and Rousseau, among others, widened and lengthened in the 19th century, the consequence of the two revolutions. If the Industrial Revolution seemed to supply all the problems frustrating the existence of a stable and humane society, the political-democratic revolution could be seen as containing many of the answers to these problems. It was the democratic revolution, especially in France, that created the vision of a political government responsible for all aspects of human society and, most important, possessed the power to wield this responsibility. This power, known as sovereignty, could be seen as holding the same relation to political science in the 19th century that capital held to economics. To a very large number of political scientists, the aim of the discipline was essentially that of analyzing the varied properties of sovereignty. There was a strong tendency on the part of such political scientists as Bentham, Austin, and Mill in England and Francis Lieber and Woodrow Wilson in the United States to see the state and its claimed sovereignty over human lives in much the same terms in which classical economists saw capitalism.

Concern with sovereignty

Among political scientists there was the same historical-evolutionary dissent from this view, however, that existed in economics. Such writers as Sir Henry Maine in England, Numa Fustel de Coulanges in France, and Otto von Gierke in Germany declared that state and sovereignty were not timeless and universal nor the results of some "social contract" envisaged by such philosophers as Locke and Rousseau but, rather, structures formed slowly through developmental or historical processes. Hence the strong interest, especially in the late 19th century, in the origins of political institutions in kinship, village, and caste, and in the successive stages of development that have characterized these institutions. In political science, as in economics, in short, the classical analytical approach was strongly rivalled by the evolutionary. Both approaches go back to the 18th century in their fundamental elements, but what is seen in the 19th century is the greater systematization and the much wider range of data employed.

Cultural anthropology. In the 19th century, anthropology also attained clear identity as a discipline. Strictly defined as "the science of man," it could be seen as superseding other specialized disciplines such as economics and political science. In practice and from the beginning, however, anthropology concerned itself overwhelmingly with primitive man. On the one hand was physical anthropology, concerned chiefly with the evolution of man as a biological species, with the successive forms and protoforms of the species, and with genetic systems such as stocks and races in the world. On the other hand was social and cultural anthropology: here the interest was in the full range of man's institutions but confined to those found in fact among existing preliterate or "primitive" peoples in Africa, Oceania, Asia, and the Americas. Above all other concepts, "culture" was the central element of this great area of anthropology, or ethnology, as it was often called to distinguish it from physical anthropology. Culture, as a concept, called attention to the nonbiological, nonracial, noninstinctual basis of the greater part of what one calls civilization: its values, techniques, ideas in all spheres. Culture, as defined in Tylor's landmark work of 1871, *Primitive Culture,* is the part of man's behaviour that is learned. From cultural anthropology more than from any other single social science has come the emphasis on the cultural foundations of man's behaviour and thought in society.

Anthropological focus on primitive man and evolutionism

Scarcely less than political science or economics, cultural anthropology shared in the themes of the two revolutions and their impact on the world. If the data that cultural anthropologists actually worked with were generally in the remote areas of the world, it was the effects of the two revolutions that, in a sense, kept opening up these parts of the world to more and more systematic inquiry. And, as was true of the other social sciences, the cultural anthropologists were immersed in problems of economics, polity, social class, and community, albeit among preliterate rather than "modern" peoples.

Overwhelmingly, without major exception indeed, the science of cultural anthropology was evolutionary in thrust in the 19th century. Edward B. Tylor and Sir John Lubbock in England, Lewis Henry Morgan in the United States, Adolf Bastian and Theodor Waitz in Germany, and all others in the main line of the study of primitive culture saw existing native societies in the world as prototypes of their own "primitive ancestors," fossilized remains, so to speak, of stages of development that western Europe had once gone through. Despite the vast array of data compiled on non-Western cultures, the same basic European-centred objectives are to be found among cultural anthropologists as among other social scientists in the century. Almost universally, then, the modern West was regarded as the latest point in a line of progress that was single and unilinear and on which all other peoples in the world could be fitted as illustrations, as it were, of Western man's own past.

Sociology. Sociology came into being in precisely these terms, and during much of the century it was not easy to distinguish between a great deal of so-called sociology and social or cultural anthropology. Even if almost no sociologists in the century made empirical studies of primitive peoples, as did the anthropologists, their interest in the origin, development, and probable future of mankind was not less great than what could be found in the writings of the anthropologists. It was Auguste Comte who coined the word sociology, and he used it to refer to what he imagined would be a single, all-encompassing, science of society that would take its place at the top of the hierarchy of sciences—a hierarchy that Comte saw as including astronomy (the oldest of the sciences historically) at the bottom and with physics, chemistry, and biology rising in that order to sociology, the latest and grandest of the sciences. There was no thought in Comte's mind—nor was there in the mind of Herbert Spencer, whose general view of sociology was very much like Comte's—of there being other, competing social sciences. Sociology would be to

The "grand" view of sociology

the whole of the social world what each of the other great sciences was to its appropriate sphere of reality.

Both Comte and Spencer believed that civilization as a whole was the proper subject of sociology. Their works were concerned, for the most part, with describing the origins and development of civilization and also of each of its major institutions. Both declared sociology's main divisions to be "statics" and "dynamics," the former concerned with processes of order in society, the latter with processes of evolutionary change in society. Both men also saw all existing societies in the world as reflective of the successive stages through which Western society had advanced in time over a period of tens of thousands of years.

Not all sociologists in the 19th century conceived their discipline in this light, however. Side by side with the "grand" view represented by Comte and Spencer were those in the century who were primarily interested in the social problems that they saw around them—consequences, as they interpreted them, of the two revolutions, the industrial and democratic. Thus in France just after midcentury, Frédéric Le Play published a monumental study of the social aspects of the working classes in Europe, *Les Ouvriers européens,* which compared families and communities in all parts of Europe and even other parts of the world. Alexis de Tocqueville, especially in the second volume of his *Democracy in America* (1835), provided an account of the customs, social structures, and institutions in America, dealing with these—and also with the social and psychological problems of Americans in that day—as aspects of the impact of the democratic and industrial revolutions upon traditional society.

The "problems" view of sociology

At the very end of the 19th century, in both France and Germany, there appeared some of the works in sociology that were to prove most lasting in their effects upon 20th-century sociology. Ferdinand Tönnies, in his *Gemeinschaft und Gesellschaft* (1887; translated as *Community and Society*), sought to explain all major social problems in the West as the consequence of the West's historical transition from the communal, status-based, concentric society of the Middle Ages to the more individualistic, impersonal, and large-scale society of the democratic-industrial period. In general terms, allowing for individual variations of theme, these were the views of Max Weber, Georg Simmel, and Émile Durkheim (all of whom also wrote in the late 19th and early 20th century). These were the men who, starting from the problems of Western society that could be traced to the effects of the two revolutions, did the most to establish the discipline of sociology as it is found for the most part in the 20th century.

Social psychology. Social psychology as a distinct discipline also originated in the 19th century, although its outlines were perhaps somewhat less clear than was true of the other social sciences. The close relation of the human mind to the social order, its dependence upon education and other forms of socialization, was well known in the 18th century. In the 19th century, however, an ever more systematic discipline came into being to uncover the social and cultural roots of human psychology and also the several types of "collective mind" that analysis of different cultures and societies in the world might reveal. In Germany, Moritz Lazarus and Wilhelm Wundt sought to fuse the study of psychological phenomena with analyses of whole cultures. Folk psychology, as it was called, did not, however, last very long in scientific esteem.

Much more esteemed, and closer to 20th-century conceptions of social psychology, were the works of such men as Gabriel Tarde, Gustave Le Bon, Lucien Lévy-Bruhl, and Émile Durkheim in France and Georg Simmel in Germany (all of whom also wrote in the early 20th century). Here, in concrete, often highly empirical studies of small groups, associations, crowds, and other aggregates (rather than in the main line of psychology during the century, which tended to be sheer philosophy at one extreme and a variant of physiology at the other) are to be found the real beginnings of social psychology. Although the point of departure in each of the studies was the nature of association, they dealt, in one degree or other, with the internal processes of psychosocial interaction, the operation of attitudes and judgments, and the social basis of personality and thought—in short, with those phenomena that would, in the 20th century, be the substance of social psychology as a formal discipline.

Early group studies

Social statistics and social geography. Two final manifestations of the social sciences in the 19th century are social statistics and social (or human) geography. At that time, neither achieved the notability and acceptance in colleges and universities that such fields as political science and economics did. Both, however, were as clearly visible by the latter part of the century as any of the other social sciences. And both were to exert a great deal of influence on the other social sciences by the beginning of the 20th century: social statistics on sociology and social psychology pre-eminently; social geography on political science, economics, history, and certain areas of anthropology, especially those areas dealing with the dispersion of races and the diffusion of cultural elements. In social statistics the key figure of the century was a Belgian, Adolphe Quetelet, who was the first, on any systematic basis, to call attention to the kinds of structured behaviour that could be observed and identified only through statistical means. It was Quetelet who brought into prominence the momentous concept of "the average man" and his behaviour. The two major figures in social or human geography in the century were Friedrich Ratzel in Germany and Paul Vidal de la Blache in France. Both broke completely with the crude environmentalism of earlier centuries, which had sought to show how topography and climate actually determine human behaviour, and they substituted the more subtle and sophisticated insights into the relationships of land, sea, and climate on the one hand and, on the other, the varied types of culture and human association that are to be found on earth.

In summary, by the end of the 19th century all the major social sciences had achieved a distinctiveness, an importance widely recognized, and were, especially in the cases of economics and political science, fully accepted as disciplines in the universities. Most important, they were generally accepted as sciences in their own right rather than as minions of philosophy.

THE 20TH CENTURY

What is seen in the 20th century is not only an intensification and spread of earlier tendencies in the social sciences but also the development of many new tendencies that, in the aggregate, make the 19th century seem by comparison one of quiet unity and simplicity in the social sciences.

In the 20th century, the processes first generated by the democratic and industrial revolutions have gone on virtually unchecked in Western society, penetrating more and more spheres of once traditional morality and culture, leaving their impress on more and more nations, regions, and localities. Equally important, perhaps in the long run far more so, is the spread of these revolutionary processes to the non-Western areas of the world. The impact of industrialism, technology, secularism, and individualism upon peoples long accustomed to the ancient unities of tribe, local community, agriculture, and religion was first to be seen in the context of colonialism, an outgrowth of nationalism and capitalism in the West. The relations of the West to non-Western parts of the world, the whole phenomenon of the "new nations," are vital aspects of the social sciences.

So too are certain other consequences, or lineal episodes, of the two revolutions. The 20th century is the century of nationalism, mass democracy, and large-scale industrialism beyond reach of any 19th-century imagination so far as magnitude is concerned. It is the century of mass warfare, of two world wars with toll in lives and property greater perhaps than the sum total of all preceding wars in history. It is the century too of totalitarianism: Communist, Fascist, and Nazi; and of techniques of terrorism that, if not novel, are to be seen on a scale and with an intensity of scientific application that could scarcely have been predicted by those who considered science and technology as unqualifiedly humane in possibility. It is a century of affluence in the West, without precedent for the masses of people, to be seen in a constantly rising standard of living and a constantly rising level of expectations.

The last is important. A great deal of the turbulence in the 20th century—political, economic, and social—is the result of desires and aspirations that have been constantly escalating and that have been passing from the white people in the West to ethnic and racial minorities among them and, then, to whole continents elsewhere. Of all manifestations of revolution, the revolution of rising expectations is perhaps the most powerful in its consequences. For, once this revolution gets under way, each fresh victory in the struggle for rights, freedom, and security tends to magnify the importance of what has not been won.

Once it was thought that, by solving the fundamental problems of production and large-scale organization, man could ameliorate other problems, those of a social, moral, and psychological nature. What in fact occurred, on the testimony of a great deal of the most notable thought and writing, was a heightening of such problems. It would appear that as man satisfies, relatively at least, the lower order needs of food and shelter, his higher order needs for purpose and meaning in life become ever more imperious.

Effects of alienation and social isolation

Thus such philosophers of history as Arnold Toynbee, Pitirim Sorokin, and Oswald Spengler have dealt with problems of purpose and meaning in history with a degree of learning and intensity of spirit not seen perhaps since St. Augustine wrote his monumental *The City of God* in the early 5th century when signs of the disintegration of Roman civilization were becoming overwhelming in their message to so many of that day. In the 20th century, though the idea of progress has certainly not disappeared, it has been rivalled by ideas of cyclical change and of degeneration of society. It is hard to miss the currency of ideas in modern times—status, community, purpose, moral integration, on the one hand, and alienation, anomie, disintegration, breakdown on the other—that reveal only too clearly the divided nature of man's spirit, the unease of his mind.

There is to be seen too, especially during later decades of the century, a questioning of the role of reason in human affairs—a questioning that stands in stark contrast with the ascendancy of rationalism in the two or three centuries preceding. Doctrines and philosophies stressing the inadequacy of reason, the subjective character of human commitment, and the primacy of faith have rivalled—some would say conquered—doctrines and philosophies descended from the Age of Reason. Existentialism, with its emphasis on the basic loneliness of the individual, on the impossibility of finding truth through intellectual decision, and on the irredeemably personal, subjective character of man's life, has proved to be a very influential philosophy in the writings of the 20th century. Freedom, far from being the essence of hope and joy, is the source of man's dread of the universe and of his anxiety for himself. Søren Kierkegaard's 19th-century intimations of anguished isolation as the perennial lot of the individual have had rich expression in the philosophy and literature of the 20th century.

It might be thought that such intimations and presentiments as these have little to do with the social sciences. This is true in the direct sense perhaps but not true when one examines the matter in terms of contexts and ambiences. The "lost individual" has been of as much concern to the social sciences as to philosophy and literature. Ideas of alienation, anomie, identity crisis, and estrangement from norms are rife among the social sciences, particularly, of course, those most directly concerned with the nature of the social bond, such as sociology, social psychology, and political science. In countless ways, interest in the loss of community, in the search for community, and in the individual's relation to society and morality have had expression in the work of the social sciences. Between the larger interests of a culture and the social sciences there is never a wide gulf—only different ways of defining and approaching these interests.

Marxist influences. The influence of Marxism in the 20th century must not be missed. Currently the works of Lenin have outstripped the Bible in distribution in the world. For hundreds of millions of persons today the ideas of Marx, as communicated by Lenin, have profound moral, even religious, significance. But even in those parts of the world, the West foremost, where Communism has exerted little direct political impact, Marxism remains a potent source of ideas. Not a few of the central concepts of social stratification and the location and diffusion of power in the social sciences come straight from Marx's insights. Far more was this the case in the Communist countries—the former Soviet Union, other eastern European countries, China, and even Asian countries in which no Communist domination exists. In all these countries, Marx's name is virtually sacrosanct. There is not the same degree of differentiation of social sciences in these countries that is found in the West. As an example, sociology hardly exists as a recognized discipline in these countries, and, by the standards of the West, the other social sciences have little more than a rather rudimentary existence. Economics alone tends to be favoured, and this is, of course, largely Marxian economics—the economics of Marx's *Das Kapital.*

But, though Marxism has had relatively little direct impact on the social sciences as disciplines in the West, it has had enormous influence on states of mind that are closely associated with the social sciences. Especially was this true during the 1930s, the decade of the Great Depression. Today signs are not lacking of a strong revival of interest in Marx that could well, through sheer numbers of its adherents, affect the nature of the social sciences in the years ahead. Socialism remains for many an evocative symbol and creed. Marx remains a formidable name among intellectuals and is still, without any question, the principal intellectual source of radical movements in politics. Such a position cannot help but influence the contexts of even the most abstract of the social sciences.

Triumph of social planning and Keynesianism

What Marx's ideas have suggested above all else in a positive way is the possibility of a society directed not by blind forces of competition and struggle among economic elements but instead by directed planning. This hope, this image, has proved a dominant one in the 20th century even where the influence of Marx and of Socialism has been at best small and indirect. It is this profound interest in central planning and governance that has given almost historic significance to the ideas of the English economist J.M. Keynes. What is called Keynesianism has as its intellectual base a very complex modification of the classical doctrines of economics—one set forth in Keynes's famous *The General Theory of Employment, Interest and Money,* published in 1935–36. Of greater influence today, however, than the strictly theoretical content of this general theory is the political impact that Keynesian ideas have had on Western democracies. For out of these ideas came the clear policy of governments dealing directly with the business cycle, of pumping money and credit into an economic system when the cycle threatens to turn downward, and of then lessening this infusion when the cycle moves upward. Above all other names in the West, that of Keynes has become identified with such policy in the democracies and with the general movement of central governments toward ever more active and constant regulation of processes once thought best left to what the classical economists thought of as natural laws. True, the root ideas of the classical economists are found in modified form even today in the works of such economists as the American Milton Friedman. But it would not be unfair to say that Keynes's name has become associated with democratic economic planning and direction in much the way that Marx's name is associated with Communist economic policies.

Freudian influences. In the general area of personality, mind, and character, the writings of Sigmund Freud have had influence on 20th-century culture and thought scarcely less than Marx's. His basic theories of the role of the unconscious mind, of the lasting effects of infantile sexuality, and of the Oedipus complex have gone beyond the discipline of psychoanalysis and even the larger area of psychiatry to areas of several of the social sciences. Anthropologists have applied Freudian concepts to their studies of primitive cultures, seeking to assess comparatively the universality of states of the unconscious that Freud and his followers held to lie in the whole human race. Some political scientists have used Freudian ideas to

illuminate the nature of authority generally, and political power specifically, seeing in totalitarianism, for example, the thrust of a craving for the security that total power can give. Sociology and social psychology have been influenced by Freudian ideas in their studies of social interaction and motivation. From Freud came the fruitful perspective that sees social behaviour and attitudes as generated not merely by the external situation but also by internal emotional needs springing from childhood—needs for recognition, authority, self-expression. Whatever may be the place directly occupied by Freud's ideas in the social sciences today, his influence upon 20th-century thought and culture generally, not excluding the social sciences, has been hardly less than Marx's.

Specialization and cross-disciplinary approaches. A major point to make about the social sciences of the 20th century is the vast increase in the number of social scientists involved, in the number of academic and other centres of teaching and research in the social sciences, and in their degree of both comprehensiveness and specialization. The explosion of the sciences generally in the 20th century—an explosion responsible for the fact that a majority of all scientists who have ever lived in human history are now alive—has had, as one of its signal elements, the explosion of the social sciences. Not only has there been development and proliferation but there has also been a spectacular diffusion of the social sciences. Beginning in a few places in western Europe and the United States in the 19th century, the social sciences, as bodies of ongoing research and centres of teaching, are today to be found almost everywhere in the world. In considerable part this has followed the spread of universities from the West to other parts of the world and, within universities, the very definite shift away from the hegemony once held by humanities alone to the near-hegemony held today by the sciences, physical and social.

Diffusion of the social sciences

Specialization has been as notable a tendency in the social sciences as in the biological and physical sciences. This is reflected not only in varieties of research but also in course offerings in academic departments. Whereas not very many years ago, a couple of dozen advanced courses in a social science reflected the specialization and diversity of the discipline even in major universities with graduate schools, today a hundred such courses are found to be not enough.

Side by side with this strong trend toward specialization, however, is another, countering trend: that of cross-fertilization and interdisciplinary cooperation. At the beginning of the century, down in fact until World War II, the several disciplines existed each in a kind of splendid isolation from the others. That historians and sociologists, for example, might ever work together in curricula and research projects would have been scarcely conceivable prior to about 1945. Each social science tended to follow the course that emerged in the 19th century: to be confined to a single, distinguishable, if artificial, area of social reality. Today, evidences are all around of cross-disciplinary work and of fusion within a single social science of elements drawn from other social sciences. Thus there are such vital areas of work as political sociology, economic anthropology, psychology of voting, and industrial sociology. Single concepts such as "structure," "function," "alienation," and "motivation" can be seen employed variously to useful effect in several social sciences. The techniques of one social science can be seen consciously incorporated into another or into several social sciences. If history has provided much in the way of perspective to sociology or anthropology, each of these two has provided perspective, and also whole techniques, such as statistics and survey, to history. In short, specialization is by no means without some degree at least of countertendencies such as fusion and synthesis.

Professionalization

Another outstanding characteristic of each of the social sciences in the 20th century is its professionalization. Without exception, the social sciences have become bodies of not merely research and teaching but also practice, in the sense that this word has in medicine or engineering. Down until about World War II, it was a rare sociologist or political scientist or anthropologist who was not a holder of academic position. There were economists and psychologists to be found in banks, industries, government, even in private consultantship, but the numbers were relatively tiny. Overwhelmingly the social sciences had visibility alone as academic disciplines, concerned essentially with teaching and with more or less basic, individual research. All this has changed profoundly, and on a vast scale, during the past three decades. Today there are as many economists and psychologists outside academic departments as within, if not more. The number of sociologists, political scientists, and demographers to be found in government, industry, and private practice rises constantly. Equally important is the changed conception or image of the social sciences. Today, to a degree unknown before World War II, the social sciences are conceived as policy-making disciplines, concerned with matters of national welfare in their professional capacities in just as sure a sense as any of the physical sciences. Inevitably, tensions have arisen within the social sciences as the result of processes of professionalization. Those persons who are primarily academic can all too easily feel that those who are primarily professional have different and competing identifications of themselves and their disciplines.

Nature of the research. The emphasis upon research in the social sciences has become almost transcending within recent decades. This situation is not at all different from that which prevails in the physical sciences and the professions in this age. Prior to about 1945, the functions of teaching and research had approximately equal value in many universities and colleges. The idea of a social (or physical) scientist appointed to an academic institution for research alone, or with research preponderant, was scarcely known. Research bureaus and institutes in the social sciences were very few and did not rival traditional academic departments and colleges as prestige-bearing entities. All of that was changed decisively beginning with the period just after World War II. From governments and foundations, large sums of money passed into the universities—usually not to the universities as such, but rather to individuals or small groups of individuals, each eminent for research. Research became the uppermost value in the social sciences (as in the physical) and hence, of course, in the universities themselves.

Use of mathematics and computers

Probably the greatest single change in the social sciences during the past generation has been the widespread introduction of mathematical and other quantitative methods. Without question, economics is the discipline in which the most spectacular changes of this kind have taken place. So great is the dominance of mathematical techniques here—resulting in the eruption of what is called econometrics to a commanding position in the discipline—that, to the outsider, economics today almost appears to be a branch of mathematics. But in sociology, political science, social psychology, and anthropology, the impact of quantitative methods, above all, of statistics, has also been notable. No longer does statistics stand alone, a separate discipline, as it did in effect during the 19th century. This area today is inseparable from each of the social sciences, though, in the field of mathematics, statistics still remains eminently distinguishable, the focus of highly specialized research and theory.

Within the past decade or two, the use of computers and of all the complex techniques associated with computers has become a staple of social-science research and teaching. Through the data storage and data retrieval of electronic computers, working with amounts and diversity of data that would call for the combined efforts of hundreds, even thousands of technicians, the social sciences have been able to deal with both the extensive and intensive aspects of human behaviour in ways that would once have been inconceivable. The so-called computer revolution in modern thought has been, in short, as vivid a phase of the social as the physical sciences, not to mention other areas of modern life. The problem as it is stated by mature social scientists is to use computers in ways in which they are best fitted but without falling into the fallacy that they can alone guide, direct, and supply vital perspective in the study of man.

Closely related to mathematical, computer, and other

quantitative aspects of the social sciences is the vast increase in the empiricism of modern social science. Never in history has so much in the way of data been collected, examined, classified, and brought to the uses of social theory and social policy alike. What has been called the triumph of the fact is nowhere more visible than in the social sciences. Without question, this massive empiricism has been valuable, indispensable indeed, to those seeking explanations of social structures and processes. Empiricism, however, like quantitative method, is not enough in itself. Unless related to hypothesis, theory, or conclusion, it is sterile, and most of the leading social scientists of today reflect this view in their works. Too many, however, deal with the gathering and classifying of data as though these were themselves sufficient.

Emphasis on empirical evidence and data gathering

It is the quest for data, for detailed, factual knowledge of human beliefs, opinions, and attitudes, as well as patterns and styles of life—familial, occupational, political, religious, and so on—that has made the use of surveys and polls another of the major tendencies in the social sciences of this century. The poll data one sees in his newspaper are hardly more than the exposed portion of an iceberg. Literally thousands of polls, questionnaires, and surveys are going on at any given moment today in the social sciences. The survey or polling method ranks with the quantitative indeed in popularity in the social sciences, both being, obviously, indispensable tools of the empiricism just mentioned.

Theoretical modes. It is not the case, however, that interest in theory is a casualty of the 20th-century fascination with method and fact. Though there is a great deal less of that grand or comprehensive theory that was a hallmark of 19th-century social philosophy and social science, there are still those persons occasionally to be found today who are engrossed in search for master principles, for general and unified theory that will assimilate all the lesser and more specialized types of theory. But their efforts and results are not regarded as successful by the vast majority of social scientists. Theory, at its best, today tends to be specific theory—related to one or other of the major divisions of research within each of the social sciences. The theory of the firm in economics, of deviance in sociology, of communication in political science, of attitude formation in social psychology, of divergent development in cultural anthropology are all examples of theory in every proper sense of the word. But each is, clearly, specific. If there is a single social science in which a more or less unified theory exists, with reference to the whole of the discipline, it is economics. Even here, however, unified, general theory does not have the sovereign sweep it had in the classical tradition of Ricardo and his followers before the true complexities of economic behaviour had become revealed.

Specific versus grand theory

Developmentalism. Developmentalism is another overall influence upon the work of the social sciences, especially within the past three decades. As noted above, an interest in social evolution was one of the major aspects of the social sciences throughout the 19th century in western Europe. In the early 20th century, however, this interest, in its larger and more visible manifestations, seemed to terminate. There was a widespread reaction against the idea of unilinear sequences of stages, deemed by the 19th-century social evolutionists to be universal for all mankind in all places. Criticism of social evolution in this broad sense was a marked element of all the social sciences, preeminently in anthropology but in the others as well. There were numerous demonstrations of the inadequacy of unilinear descriptions of change when it came to accounting for what actually happened, so far as records and other evidences suggested, in the different areas and cultures of the world.

Influences of national growth and development

Beginning in the late 1940s and the 1950s, however, there was a resurgence of developmental ideas in all the social sciences—particularly with respect to studies of the new nations and cultures that were coming into existence in considerable numbers. Studies of economic growth and of political and social development have become more and more numerous. Although it would be erroneous to see these developmental studies as simple repetitions of those of the 19th-century social evolutionists, there are, nevertheless, common elements of thought, including the idea of stages of growth and of change conceived as continuous and cumulative and even as moving toward some more or less common end. At their best, these studies of growth and development in the new nations, by their counterposing of traditional and modern ways, tell a good deal about specific mechanisms of change, the result of the impact of the West upon outlying parts of the world. But as more and more social scientists have recently become aware, efforts to place these concrete mechanisms of change into larger, more systematic models of development all too commonly succumb to the same faults of unilinearity and specious universalism that early-20th-century critics found in 19th-century social evolution.

Social-systems approach. Still another major tendency in all of the social sciences since World War II has been the interest in "social systems." The behaviour of individuals and groups is seen as falling into multiple interdependencies, and these interdependencies are considered sufficiently unified to warrant use of the word "system." Although there are clear uses of biological models and concepts in social-systems work, it may be fair to say that the greatest single impetus to development of this area was widening interest after World War II in cybernetics—the study of human control functions and of the electrical and mechanical systems that could be devised to replace or reinforce them. Concepts drawn from mechanical and electrical engineering have been rather widespread in the study of social systems.

In social-systems studies, the actions and reactions of individuals, or even of groups as large as nations, are seen as falling within certain definable, more or less universal patterns of equilibrium and disequilibrium. The interdependence of roles, norms, and functions is regarded as fundamental in all types of group behaviour, large and small. Each social system, as encountered in social-science studies, is a kind of "ideal type," not identical to any specific "real" condition but sufficiently universal in terms of its central elements to permit useful generalization.

Structuralism and functionalism. Structuralism in the social sciences is closely related to the theory of the social system. Although there is nothing new about the root concepts of structuralism—they may be seen in one form or other throughout Western thought—there is no question but that in the present century this view of behaviour has become a dominant one in many fields. At bottom it is a reaction against all tendencies to deal with human thought and behaviour atomistically—that is, in terms of simple, discrete units of either thought, perception, or overt behaviour. In psychology, structuralism in its oldest sense simply declares that perception occurs, with learning following, in terms of experiences or sensations in various combinations, in discernible patterns or gestalten. In sociology, political science, and anthropology, the idea of structure similarly refers to the repetitive patternings that are found in the study of social, economic, political, and cultural existence. The structuralist contends that no element can be examined or explained outside its context or the pattern or structure of which it is a part. Indeed, it is the patterns, not the elements, that are the only valid objects of study.

Emphasis on pattern and interdependence

What is called functionalism in the social sciences today is closely related to structuralism, with the term structural-functional a common one, especially in sociology and anthropology. Function refers to the way in which behaviour takes on significance, not as a discrete act but as the dynamic aspect of some structure. Biological analogies are common in theories of structure and function in the social sciences. Very common is the image of the biological organ, with its close interdependence to other organs (as the heart to the lung) and the interdependence of activities (as circulation to respiration).

Interactionism. Interaction is still another concept that has had wide currency in the social sciences of the 20th century. Social interaction—or, as it is sometimes called, symbolic interaction—refers to the fact that the relationships among two or more groups or human beings are never one-sided, purely physical, or direct. Always there is reciprocal influence, a mutual sense of "otherness." And

always the presence of the "other" has crucial effect in one's definition of not merely what is external but what is internal. One acquires one's individual sense of identity from interactions with others beginning in infancy. It is the initial sense of the other person—mother, for example—that in time gives the child its sense of self, a sense that requires continuous development through later interactions with others. From the point of view of interactionist theory, all one's perceptions of and reactions to the external world are mediated or influenced by prior ideas, valuations, and assessments. Always one is engaged in socialization or the modification of one's mind, role, and behaviour through contact with others. (R.A.N./Ed.)

Anthropology

Anthropology is "the science of humanity," which studies human beings in aspects ranging from the biology and evolutionary history of *Homo sapiens* to the features of society and culture that decisively distinguish humans from other animal species. Because of the diverse subject matter it encompasses, anthropology has become, especially since the middle of the 20th century, a collection of more specialized fields. Physical anthropology is the branch that concentrates on the biology and evolution of humanity. It is discussed in greater detail in the article EVOLUTION, HUMAN. The branches that study the social and cultural constructions of human groups are variously recognized as belonging to cultural anthropology (or ethnology), social anthropology, linguistic anthropology, and psychological anthropology (see below). Archaeology, as the method of investigation of material remains—mostly those of prehistoric cultures—has been an integral part of anthropology since it became a self-conscious discipline in the latter half of the 19th century. (For further discussion of archaeology, see HISTORY, STUDY OF: *Ancillary fields.*)

A collection of specialized fields

OVERVIEW

Throughout its existence as an academic discipline, anthropology has been located at the intersection of natural science and humanities. The biological evolution of *Homo sapiens* and the evolution of the capacity for culture that distinguishes humans from all other species are indistinguishable from one another. While the evolution of the human species is a biological development like the processes that gave rise to the other species, the historical appearance of the capacity for culture initiates a qualitative departure from other forms of adaptation, based on an extraordinarily variable creativity not directly linked to survival and ecological adaptation. The historical patterns and processes associated with culture as a medium for growth and change, and the diversification and convergence of cultures through history, are thus major foci of anthropological research.

In the middle of the 20th century, the distinct fields of research that separated anthropologists into specialties were (1) physical anthropology, emphasizing the biological process and endowment that distinguishes *Homo sapiens* from other species, (2) archaeology, based on the physical remnants of past cultures and former conditions of contemporary cultures, usually found buried in the earth, (3) linguistic anthropology, emphasizing the unique human capacity to communicate through articulate speech and the diverse languages of humankind, and (4) social and/or cultural anthropology, emphasizing the cultural systems that distinguish human societies from one another, and the patterns of social organization associated with these systems. By the middle of the 20th century, many American universities also included (5) psychological anthropology, emphasizing the relationships among culture, social structure, and the human being as a person.

The concept of culture as the entire way of life or system of meaning for a human community was a specialized idea shared mainly by anthropologists until the latter half of the 20th century. However, it had become a commonplace by the beginning of the 21st century. The study of anthropology as an academic subject had expanded steadily through those 50 years, and the number of professional anthropologists had increased with it. The range and specificity of anthropological research and the involvement of anthropologists in work outside of academic life have also grown, leading to the existence of many specialized fields within the discipline. Theoretical diversity has been a feature of anthropology since it began and, although the conception of the discipline as "the science of humanity" has persisted, some anthropologists now question whether it is possible to bridge the gap between the natural sciences and the humanities. Others argue that new integrative approaches to the complexities of human being and becoming will emerge from new subfields dealing with such subjects as health and illness, ecology and environment, and other areas of human life that do not yield easily to the distinction between "nature" and "culture" or "body" and "mind."

Anthropology in 1950 was—for historical and economic reasons—instituted as a discipline mainly in western Europe and North America. Field research was established as the hallmark of all the branches of anthropology. While some anthropologists studied the "folk" traditions in Europe and America, most were concerned with documenting how people lived in nonindustrial settings outside these areas. These finely detailed studies of everyday life of people in a broad range of social, cultural, historical, and material circumstances were among the major accomplishments of anthropologists in the second half of the 20th century.

Beginning in the 1930s, and especially in the post-World War II period, anthropology was established in a number of countries outside western Europe and North America. Very influential work in anthropology originates in Japan, India, China, Mexico, Brazil, Peru, South Africa, Nigeria, and several other Asian, Latin American, and African countries. The world scope of anthropology, together with the dramatic expansion of social and cultural phenomena that transcend national and cultural boundaries, has led to a shift in anthropological work in North America and Europe. Research by Western anthropologists is increasingly focused on their own societies, and there have been some studies of Western societies by non-Western anthropologists. By the end of the 20th century, anthropology was beginning to be transformed from a Western—and some have said "colonial"—scholarly enterprise into one in which Western perspectives are regularly challenged by non-Western ones. (R.W.N.)

HISTORY OF ANTHROPOLOGY

The modern discourse of anthropology crystallized in the 1860s, fired by advances in biology, philology, and prehistoric archaeology. In *The Origin of Species* (1859), Charles Darwin affirmed that all forms of life share a common ancestry. Fossils began to be reliably associated with particular geological strata, and fossils of recent human ancestors were discovered, most famously the first Neanderthal specimen, unearthed in 1856. In 1871 Darwin published *The Descent of Man,* which argued that human beings shared a recent common ancestor with the great African apes. He identified the defining characteristic of the human species as a relatively large brain size and deduced that the evolutionary advantage of the human species was intelligence, which yielded language and technology.

E.B. Tylor

The pioneering anthropologist E.B. Tylor concluded that as intelligence increased so civilization advanced. All past and present societies could be arranged in an evolutionary sequence. Archaeological findings were organized in a single universal series (Stone Age, Iron Age, Bronze Age, etc.) thought to correspond to stages of economic organization from hunting and gathering to pastoralism, agriculture, and industry. Some contemporary peoples (hunter-gatherers, such as the Australian Aboriginals and the Kalahari San, or pastoralists such as the Bedouin) were regarded as "primitive," laggards in evolutionary terms, representing stages of evolution through which all other societies had passed. They bore witness to early stages of human development, while the industrial societies of northern Europe and the United States represented the pinnacle of human achievement.

Darwin's arguments were drawn upon to underwrite the universal history of the Enlightenment, according to which the progress of human institutions was inevitable, guaranteed by the development of rationality. It was assumed that

technological progress was constant and that it was matched by developments in the understanding of the world and in social forms. Tylor advanced the view that all religions had a common origin, in the belief in spirits. The original religious rite was sacrifice, which was a way of feeding these spirits. Modern religions retained some of these primitive features, but as human beings became more intelligent, and so more rational, primitive superstitions were gradually refined and would eventually be abandoned. James George Frazer posited a progressive and universal progress from faith in magic through to belief in religion and, finally, to the understanding of science.

J.F. McLennan, Lewis Henry Morgan, and other writers argued that there was a parallel development of social institutions. The first humans were promiscuous (like, it was thought, the African apes), but at some stage blood ties were recognized between mother and children and incest between mother and son was forbidden. In time more restrictive forms of mating were introduced and paternity was recognized. Blood ties began to be distinguished from territorial relationships, and distinctive political structures developed beyond the family circle. At last monogamous marriage evolved. Paralleling these developments, technological advances produced increasing wealth, and arrangements guaranteeing property ownership and regulating inheritance became more significant. Eventually, the modern institutions of private property and territorially based political systems developed, together with the nuclear family.

An alternative to this Anglo-American "evolutionist" anthropology established itself in the German-speaking countries. Its scientific roots were in geography and philology, and it was concerned with the study of cultural traditions and with adaptations to local ecological constraints rather than with universal human histories. This more particularistic and historical approach was diffused to the United States at the end of the 19th century by the German-trained scholar Franz Boas. Skeptical of evolutionist generalizations, Boas advocated instead a "diffusionist" approach. Rather than graduating through a fixed series of intellectual, moral, and technological stages, societies or cultures changed unpredictably, as a consequence of migration and borrowing.

Fieldwork. The first generation of anthropologists had tended to rely on others—locally based missionaries, colonial administrators, and others—to collect ethnographic information, often guided by questionnaires that were issued by metropolitan theorists. In the late 19th century, several ethnographic expeditions were organized, often by museums. As reports on customs came in from these various sources, the theorists would collate the findings in comparative frameworks, to illustrate the course of evolutionary development or to trace local historical relationships.

Ethnographic expeditions

The first generation of professionally trained anthropologists began to undertake intensive fieldwork on their own account in the early 20th century. As theoretically trained investigators began to spend long periods alone in the field, on a single island or in a particular tribal community, the object of investigation shifted. The aim was no longer to establish and list traditional customs. Fieldworkers began to record the activities of flesh-and-blood human beings going about their daily business. To get this sort of material, it was no longer enough to interview local authority figures. The fieldworker had to observe people in action, off-guard, to listen to what they said to each other, to participate in their daily activities. The most famous of these early intensive ethnographic studies was carried out between 1915 and 1918 by Bronisław Malinowski in the Trobriand Islands (now Kiriwina Islands) off the southeastern coast of New Guinea, and his Trobriand monographs, published between 1922 and 1935, set new standards for ethnographic reportage.

These new field studies reflected and accelerated a change of theoretical focus from the evolutionary and historical interests of the 19th century. Inspired by the social theories of Émile Durkheim and the psychological theories of Wilhelm Wundt and others, the ultimate aim was no longer to discover the primitive origins of Western customs but rather to explain the purposes that were served by particular institutions or religious beliefs and practices. Malinowski explained that Trobriand magic was not simply poor science. The "function" of garden magic was to sustain the confidence of gardeners, whose investments could not be guaranteed. His colleague, A.R. Radcliffe-Brown, adopted a more sociological, Durkheimian line of argument, explaining, for example, that the "function" of ancestor worship was to sustain the authority of fathers and grandfathers and to back up the claims of family responsibility. Perhaps the most influential sociological explanation of "primitive" institutions was Marcel Mauss's account of gift exchanges, illustrated by such diverse practices as the "*kula* ring" cycle of exchange of the Trobriand Islanders and the potlatch of the Kwakiutl of the Northwest Pacific Coast of North America. Mauss argued that apparently irrational forms of economic consumption made sense when they were properly understood, as modes of social competition, regulated by strict and universal rules of reciprocity.

Marcel Mauss and the *kula* ring

Social and cultural anthropology. A distinctive "social" or "cultural" anthropology emerged in the 1920s. It was associated with the social sciences and linguistics, rather than with human biology and archaeology. In Britain in particular social anthropologists came to regard themselves as comparative sociologists, but the assumption persisted that anthropologists were primarily concerned with "primitive" peoples, and in practice evolutionary ways of thinking may often be discerned below the surface of functionalist argument that represents itself as ahistorical. A stream of significant monographs and comparative studies appeared in the 1930s and '40s that described and classified the social structures of what were termed tribal societies. In *African Political Systems* (1940), Meyer Fortes and Edward Evans-Pritchard proposed a triadic classification of African polities. Some African societies (*e.g.*, the San) were organized into kin-based bands. Others (*e.g.*, the Nuer and the Tallensi) were federations of unilineal descent groups, each of which was associated with a territorial segment. Finally, there were territorially based states (*e.g.*, those of the Tswana of southern Africa and the Kongo of central Africa) or the emirates of northwest Africa, in which kinship and descent regulated only domestic relationships. Kin-based bands lived by foraging, lineage-based societies were often pastoralists, and the states combined agriculture, pastoralism, and trade. In effect, this was a transformation of the evolutionist stages into a synchronic classification of types. Though speculations about origins were discouraged, it was apparent that the types could easily be rearranged in a chronological sequence from the most primitive to the most sophisticated.

There were similar attempts to classify systems of kinship and marriage, the most famous being that of the French anthropologist Claude Lévi-Strauss. In 1949 he presented a classification of marriage systems from diverse localities, again within the framework of an implicit evolutionary series. The crucial evolutionary moment was the introduction of the incest taboo, which obliged men to exchange their sisters and daughters with other men in order to acquire wives for themselves and their sons. These marriage exchanges in turn bound family groups together into societies. In societies organized by what Lévi-Strauss termed "elementary systems" of kinship and marriage, the key social units were exogamous descent groups, that is, people descended from a common ancestor who are prohibited from marrying one another. He represented the Australian Aboriginals as the most fully realized example of an elementary system, while most of the societies with complex kinship systems were to be found in the modern world, in complex civilizations.

Claude Lévi-Strauss

American anthropology since the 1950s. In the United States, a "culture-and-personality" school developed that drew rather on new movements in psychology (particularly psychoanalysis and Gestalt psychology). Later developments in the social sciences resulted in the emergence of a positivist cross-cultural project, associated with G.P. Murdock at Yale University, which applied statistical methods to a sample of world cultures and attempted to establish universal functionalist relationships between forms of marriage, descent systems, property relationships, and other

variables. Under the influence of the American social theorist Talcott Parsons, the anthropologists at Harvard University were drawn into team projects with sociologists and psychologists. They came to be regarded as the specialists in the study of "culture" within the framework of an interdisciplinary social science.

The new evolutionists

In the 1950s and '60s, evolutionist ideas gained fresh currency in American anthropology, where they were cast as a challenge to the relativism and historical particularism of the Boasians. Some of the new evolutionists (led by Leslie White) reclaimed the abandoned territory of Victorian social theory, arguing for a coherent world history of human development, through a succession of stages, from a common primitive base. The more developed a society, the more complex its organization and the more energy it consumed. White believed that energy consumption was the gauge of cultural advance. Another tendency, led by Julian Steward, argued rather for an evolutionism that was more directly Darwinian in inspiration. Cultural practices were to be treated as modes of adaptation to specific environmental challenges. More skeptical than White about traditional models of unilineal evolution, Steward urged the study of particular evolutionary processes within enduring culture areas, in which societies with a common origin were exposed to similar ecological constraints. Students of White and Steward, including Marshall Sahlins, revived classic evolutionist questions about the origins of the state and the consequences of technological progress.

The institutional development of anthropology in Europe was strongly influenced by the existence of overseas empires, and in the aftermath of World War II anthropologists were drawn into development programs in the so-called Third World. In the United States, anthropologists had traditionally studied the native peoples of North and Central America. During World War II, however, they were called upon to apply their expertise to assist the war effort, along with other social scientists. As the United States became increasingly influential in the world, in the aftermath of the war, the profession grew explosively. In the 1950s and '60s, important field studies were carried out by American ethnographers working in Indonesia, in East and West Africa, and in the many societies in the South Seas that had been brought under direct or indirect American control as a result of the war in the Pacific.

In the view of some critics, social and cultural anthropology was becoming, in effect, a Western social science that specialized in the study of colonial and postcolonial societies. The war in Vietnam fueled criticism of American engagement in the Third World and precipitated a radical shift in American anthropology. There was general disenchantment with the project of "modernizing" the new states that had emerged after World War II, and many American anthropologists began to turn away from the social sciences.

American anthropology divided between two intellectual tendencies. One school, inspired by modern developments in genetics, looked for biological determinants of human cultures and sought to revive the traditional alliance between cultural anthropology and biological anthropology. Another school insisted that cultural anthropology should aim to interpret other cultures rather than to seek laws of cultural development or cultural integration and that it should therefore seek a place among the humanities rather than with the biological sciences or the social sciences.

"Interpretive" anthropology

Clifford Geertz was the most influential proponent of an "interpretive" anthropology. This represented a movement away from biological frameworks of explanation and a rejection of sociological or psychological preoccupations. The ethnographer was to focus on symbolic communications, and so rituals and other cultural performances became the main focus of research. Sociological and psychological explanations were left to other disciplines. In the next generation, a radically relativist version of Geertz's program became influential. It was argued that cultural consensus is rare and that interpretations are therefore always partial. Cultural boundaries are provisional and uncertain, identities fragile and fabricated. Consequently ethnographers should represent a variety of discordant voices, not try to identify a supposedly normative cultural view. In short, it was an illusion that objective ethnographic studies could be produced and reliable comparisons undertaken.

European anthropology since the 1950s. In Europe the social science program remained dominant, though it was revitalized by a new concern with social history. Some European social scientists became leaders of social thought, among them sociologist Pierre Bourdieu and anthropologists Mary Douglas, Louis Dumont, Ernest Gellner, and Claude Lévi-Strauss. Elsewhere, particularly in some formerly colonial countries in Latin America, Asia, and Africa, local traditions of anthropology established themselves. While anthropologists in these countries were responsive to theoretical developments in the traditional centres of the discipline, they were also open to other intellectual currents, because they were typically engaged in debates with specialists from other fields about developments in their own countries.

The rise of ethnography

Empirical research flourished despite the theoretical diversity. Long-term fieldwork was now commonly backed up by historical investigations, and ethnography came to be regarded by many practitioners as the core activity of social and cultural anthropology. In the second half of the 20th century, the ethnographic focus of anthropologists changed decisively. The initial focus had been on "primitive" peoples. Later, ethnographers specialized in the study of Third World societies, including the complex villages and towns of Asia. Fieldwork began increasingly to be carried out in European societies and among ethnic minorities, church communities, and other groups in the United States. In the formerly colonized societies, local anthropologists began to dominate ethnographic research, and community leaders increasingly insisted on controlling the agenda of fieldworkers.

The liveliest intellectual developments were perhaps to be found beyond the mainstream. Fresh specializations emerged as legitimate, recognized fields of study, notably the anthropology of women in the 1970s, and in the following decades medical anthropology, psychological anthropology, visual anthropology, the anthropology of music and dance, and demographic anthropology. The anthropology of the 21st century is polycentric and cosmopolitan, and it is not entirely at home among the biological or social sciences or in the humanities. (A.Ku.)

CULTURAL ANTHROPOLOGY

Cultural anthropology is that major division of anthropology that explains culture in its many aspects. It is anchored in the collection, analysis, and explanation (or interpretation) of the primary data of extended ethnographic field research. This discipline, in both America and Europe, has long cast a wide net and includes various approaches. It has produced such collateral approaches as culture-and-personality studies, culture history, cultural ecology, cultural materialism, ethnohistory, and historical anthropology. These subdisciplines variously exploit methods from the sciences and the humanities. Cultural anthropology has become a family of approaches oriented by the culture concept.

The central tendencies and recurrent debates since the mid-19th century have engaged universalist versus particularist perspectives, scientific versus humanistic perspectives, and the explanatory power of biology (nature) versus that of culture (nurture). Two persistent themes have been the dynamics of culture change and the symbolic meanings at the core of culture.

Tylor's definition of culture

The definition of culture has long provoked debate. The earliest and most quoted definition is the one formulated in 1871 by E.B. Tylor: "Culture or Civilization, taken in its wide ethnographic sense, is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society." Three things of enduring relevance are to be remarked in this definition. First, it treats culture and civilization as interchangeable terms. Second, it emphasizes ethnography. And third, it singles out that which is learned by life in society rather than what is inherited biologically.

In respect to culture and civilization, Tylor collapses the distinction between the total social legacy of a human group, including every mundane matter from pot-making

to toilet practices, and their most refined attainments, such as the fine arts, that has been at the heart of the debate over what culture is. On the second point, he emphasizes what has continued to be the anchor of cultural anthropology in ethnographic fieldwork and writing, the positioning and gender of the ethnographer, and bias in ethnographic data has undergone increasingly close scrutiny. On the third point, by emphasizing what is socially learned rather than what is biologically transmitted, Tylor points up the enduring problem of distinguishing between biological and cultural influences, between nature and nurture.

Tylor's definition is taken as the inception of the awareness of culture in anthropology, but classical thinkers such as Herodotus and Tacitus were also aware of differences in beliefs and practices among the diverse peoples of the then-known world—that is, of cultural difference. It was the Age of Exploration and Discovery that exposed the breadth of human diversity, posing fundamental questions of universality and particularity in human life ways that have become the province of cultural anthropology. In the face of such diversity, Enlightenment thinkers sought to discover what could still be taken as universally reasonable—enlightened or truly civilized—in the living out of human relationships. The French Enlightenment emphasized universals grounded in human reason against which the German thinkers, most notably J.G. Herder, spoke of *Kultur*, which is to say the particular identity-defining differences characteristic of peoples and nations. This universalism/particularism debate between French and German thinkers, which is a version of the debate between Classicism and Romanticism, has continued to be central in cultural anthropology. There is also the related debate between idealism and materialism: European Idealism emphasized the subtle meaningfulness of local configurations of thought and value over against the practical focus on utilitarian analysis of health, material well-being, and survival. This idealism flourished in German anthropology in the late 19th century, notably in the work of Rudolf Virchow and Adolf Bastian, and influenced Franz Boas, a longtime professor at Columbia University, who trained most of the formative generation of 20th-century American anthropologists. The debate between idealism and materialism in cultural anthropology continues today.

J.G. Herder and *Kultur*

American cultural anthropology. The idealism of Boasian cultural anthropology found its first challenge in 19th-century cultural evolutionism, which had its origins in the early modern notion of the Great Chain of Being. Stimulated mainly by Darwinian thought, 19th-century classical evolutionism arranged the different life ways of the world on a hierarchical and unilinear ladder proceeding from savagery to barbarism to civilization, taking as exemplary of the latter such evolved civilizations as the Euro-American and the Asiatic. The second tendency in this thought was the identification of "race" with culture. One saw the "lower races," most of them with black or brown skin, as having, through biological incapacity for culture, fallen behind or lost out in the evolutionary competition for "the survival of the fittest." (For further discussion of this issue, see EVOLUTION, HUMAN: *Race.*)

These unilinear hierarchies and their presumptions were challenged by the Boasians on a number of fronts. First, their fieldwork, largely undertaken among American Indians, showed the widespread influences of diffusion between cultures, stimulating culture change that rendered any simple picture of unilinear evolution untenable. All cultures learned from each other throughout their histories. Also, the discovery that cultural adaptation to particular local physical environment had an important influence on evolution led to a more pluralistic and multilineal approach to culture change. The comparison of cultures that arose in early 20th-century anthropology produced diverse theoretical and methodological consequences, most notably the concept of cultural relativism, a theory of culture change or acculturation, and an emphasis on the study of symbolic meaning. Perhaps the most important achievement of Boas and his students was the demonstration that there is no necessary connection between culture and "race," that the capacity for culture of specific groups was not genetically controlled, and that the freedom to create cultures independent of biology was one of the great achievements of human evolution.

French theoretical contributions. French ethnology under the influence of Émile Durkheim and Marcel Mauss and their successors emphasized the study of culture, or society, as a total system with a definite "structure" consisting of elements that "functioned" both to adapt to changing circumstances and to reproduce its integral structure. The total system approach influenced British social anthropology in the form of Malinowski's functionalism and Radcliffe-Brown's attention to the dynamics of social structure. British structural-functionalism became influential, even in the United States, as a countercurrent to the cultural emphasis of American anthropology. In part this emphasis is present because, after World War II, many American anthropologists did ethnographic fieldwork in Africa, South Asia, and the Pacific, where British-trained social anthropologists were the pioneers. The emphasis on the study of whole cultures and on cultures as systems in American cultural anthropology, often called *holism*, also shows both French and British influence.

British structural-functionalism

Although it began in the study of social structures, "structuralism" aimed at understanding the universals of mental structures. It was mainly developed by Claude Lévi-Strauss, who was much influenced by Durkheim and Mauss as well as by structural linguistics. Structuralism affected American cultural anthropology, harmonizing with idealist elements and the treatment of culture as first of all patterns of belief or ideas which eventuated in practical activity. Only later, in the last several decades of the 20th century, was practice given primary emphasis in the analysis of social and political life by theorists such as Pierre Bourdieu, and the analysis of discourse by linguistic anthropologists such as Michael Silverstein. The interaction between ideas on the one hand and social and political behaviour on the other has long been a contested issue in cultural anthropology, and it remains so.

The configurational approach. The development of American cultural anthropology between the two World Wars and into the decade of the 1960s was significantly shaped by anthropological linguist Edward Sapir, who demonstrated the determinative effect of language on culture and worldview and who argued that culture is largely psychological. Since language is central to the task of the ethnographer, to learning, to the expression of thought and values, and to the transmission of culture, Sapir's language-anchored perspectives have had important and continuing resonance. His psychological emphasis was influential in the culture-and-personality movement that flourished under other Boasians, notably Margaret Mead and Ruth Benedict.

Margaret Mead and Ruth Benedict

The Boasian resistance to the sweeping and confining generalizations of classic evolutionism had two consequences: an emphasis on culture change at a specific level of analysis and a priority on studying the patterns or configurations of local cultural beliefs and values. Pattern and configuration became key concepts for explaining the relation of culture traits to each other and the study of local patterning of cultural traits and changes over time. Benedict's popular presentation, *Patterns of Culture* (1934), though espousing a cultural psychology, is an example, as is the austere and massive *Configurations of Culture Growth* (1944) by another of Boas's students, Alfred Kroeber.

This emphasis on the study of internal patterns and configurations of particular cultures as these are expressed in language led in two directions: to "cultural relativism" and to the study of "culture contact," or "acculturation." "Relativism," which resists universal judgments of any kind, is usually identified with American cultural anthropology, mainly through the work of Benedict and Melville Herskovits. It remains a persistent challenge to the generalizing impulse in anthropology and in the academy.

Cultural change and adaptation. Ethnographic fieldwork had been undertaken mainly in colonial situations characterized by contact between conquering and conquered cultures. This experience produced a theory of cultural cross-fertilization (acculturation) and culture change. A legacy of colonialism was the great differential between developed and underdeveloped parts of the world. The "de-

velopment project" undertaken by the richer nations after World War II to relieve colonial poverty and diminish global inequities has produced various cultural theories of development based on continuing anthropological research as well as strong critiques of the discipline's role in development.

Cultural anthropology has maintained its concern for the history of change in particular cultures. Kroeber was the most notable cultural historian among Boas's students, examining change over the long term on a scale that connected easily with the historical sociology of Max Weber and the social history of Fernand Braudel. The last two decades of the 20th century witnessed a striking invigoration of historical anthropology that took issue with utilitarian and materialist interpretations of cultural stability and change, emphasizing the importance of symbols and their meaning for all human action. Marshall Sahlins was a leading proponent of this school of "historical anthropology."

Marshall Sahlins

Cultural ecology also has its roots in an earlier cultural anthropology, particularly the study of the geographic and environmental context of culture change. The neo-evolutionist Leslie White reacted to the idealism of the cultural approach, turning his attention to the progress of technology in harnessing energy to serve the survival and subsistence need of cultures. Cultural ecology has sought to produce a more quantitative discipline than is characteristic of most cultural anthropology, which has remained rooted in the humanities.

Culture and the humanities. The humanistic roots of cultural anthropology produced some of the major tendencies of the latter half of the 20th century. Cultural anthropology in America has long studied the folklore, music, art, worldview, and indigenous philosophies of other cultures. Humanistic scholarship typically makes qualitative or interpretive statements about complex patternings or configurations of experience and local meaning such as can not easily be done by formal scientific procedures. In the 1950s, Kroeber and Clyde Kluckhohn, two of the most eminent anthropologists of the period, undertook a major effort to assay the meaning of "culture" in anthropology; they concluded that it was best understood as the knowledge, belief, and habits embodied in symbolic discourse. The symbolic anthropology that flourished in cultural anthropology from the 1960s to the '80s was mainly concerned with the interpretation of the complex meaning of symbols in local experience.

An important contribution to redefining cultural anthropology in the 1970s was the interpretive movement promoted by Clifford Geertz. He argued that the main consequence of fieldwork was the anthropologists' densely interwoven, symbol-laden field texts ("field notes") and their main products were the texts interpreting these texts, the ethnographies themselves. Anthropological work should thus be seen as a text-oriented interpretive task practiced on the rich complexities of culture and social action. A further step along this path challenged anthropology with the "writing culture" movement, which pointed up the biases implicit in the anthropologist's positioning in field research, and his or her choice of voices to hear and materials to write about in the ethnographic text. Geertz thus enabled many anthropologists of all persuasions to recognize the limits of objectivity, and the inevitable "partiality" of anthropological practice and publication. A related critique came from feminists in anthropology who pressed the case of culturally influenced gender bias in fieldwork and writing.

These developments were followed in the 1990s by the "writing against culture" movement, which expressed misgivings about a common form of anthropological thought that imposed excessive and disadvantaging "otherness" on the cultures and peoples studied. This movement implicitly reasserted the humanist universalism of anthropology. Other cultures were described in terms that distanced and dehumanized them. This was a very direct and forceful challenge to customary descriptive and categorizing practices, and it provoked strong debate in the discipline. The exchange between the Sri Lankan anthropologist Gananath Obeyesekere and the American anthropologist Marshall Sahlins concerning the interpretation of precolonial native thought in the Hawaiian Islands was a late 20th-century episode in the continuing debate between cultural universalism and cultural particularism.

Cultural universalism versus cultural particularism

Symbolic anthropology has given rise to a new theme, the role of metaphor, or more broadly all the tropes, or figures of speech, as symbolic representations of proper conduct. This is an ancient scholarly interest, dating from Aristotle in Western thought, but not unique to Western civilization. Partaking of both humanistic and scientific analysis, this approach is fruitful both for insight into the mind and the organization of experience, and for the understanding of the constraints and creative possibilities the "play of tropes" contributes to expressive culture.

The turn of the millennium saw a renewal of the relationship between anthropology and the humanities, as the concept of culture has been adopted as the centrepiece of "cultural studies," with its focal interest in "multiculturalism." The self-identification of many minorities in American society brought with it a large number of new areas of study in the humanities. Humanists, to be sure, were, from the turn of the last century, influenced by the anthropological work of Frazer and others. However, these new humanistic approaches to the study of the relation of changing thought and value to the changing social, political, and economic circumstances of a globalizing market, though not grounded in extended fieldwork and empirical ethnography, pose an important challenge to anthropology's claim to be the interpreter and arbiter of the culture concept. "Cultural studies" pose a challenge of collaboration between anthropology and the humanities. The recent movement away from the study of small-scale societies and a new focus on the study of emergent "public cultures" in the global arena has been a significant anthropological response to this new interest in culture in the humanities. (For further discussion of this issue, see GLOBALIZATION AND CULTURE.) (J.W.Fe.)

LINGUISTIC ANTHROPOLOGY

Linguistic anthropologists argue that human production of talk and text, made possible by the unique human capacity for language, is a fundamental mechanism through which people create culture and social life. Contemporary scholars in the discipline explore how this creation is accomplished by using many methods, but they emphasize the analysis of audio- or video-recordings of "socially-occurring" discourse—that is, talk and text that would appear in a community whether or not the anthropologist was present. This method is preferred because differences in how different communities understand the meaning of speech acts, such as "questioning," may shape results derived from investigator-imposed elicitation, such as "interviewing," in unpredictable ways.

A central question for linguistic anthropology is whether differences in culture and structural usage among diverse languages promote differences among human communities in how the world is understood. Local cultures of language may prefer certain forms of expression and avoid others. For instance, while the vocabulary of English includes an elaborate set of so-called absolute directionals (words such as *north* and *southwest*), most speakers seldom use these terms for orientation, preferring vocabulary that is relative to a local context, such as *downhill* or *left*.

Local preferences

"Cultures of language" may cross linguistic boundaries. Thus Native American Puebloans, speaking languages of four unrelated families, avoid using different languages in the same utterance even when speakers are multilingual, and do not allow everyday speech to intrude into religious contexts. By contrast, their Spanish-speaking neighbours often switch between Spanish and English and value colloquial forms in worship, as is evident in their folk masses composed in everyday language.

An important line of research explores how "cultural models"—local understandings of the world—are encoded in talk and text. Students of "language ideologies" look at local ideas about how language functions. A significant language ideology associated with the formation of modern nation-states constructs certain ways of speaking as "standard languages;" once a standard is defined, it is treated as prestigious and appropriate, while other languages or dialects are marginalized and stigmatized.

Linguistic anthropologists explore the question of how linguistic diversity is related to other kinds of human difference. Franz Boas insisted that "race," "language," and "culture" are quite independent of one another. For instance, communities of Pygmy hunters in East Africa are biologically and culturally distinct from neighbouring cultivators, but both groups share the same Bantu languages. In contrast, the Puebloan peoples of the U.S. Southwest share a common cultural repertoire, but they speak languages that belong to four different and unrelated families.

The approximately 6,000 languages spoken in the world today are divided by historical linguists into genealogical families (languages descended from a common ancestor). Some subgroups—such as the African Bantu languages within the Niger-Congo language family, which include hundreds of languages and cover enormous geographic areas—are very large. Others, such as Keresan in the U.S. Southwest, with two closely related varieties, are very small. Accounting for this difference is a significant topic of research. Geographically extensive and numerically large families may result from major technological innovations, such as the adoption of cultivation, which permit the community of innovators, and its language, to expand at the expense of neighbouring groups. An alternative possibility is that certain kinds of physical environment, such as the Eurasian steppes, favour language spread and differentiation, whereas other kinds of physical environments, such as the mountainous zones, favour the proliferation of small linguistic communities, regardless of technology.

The question of why one language expands and diversifies at the expense of its neighbours was particularly acute at the beginning of the 21st century, when a few world languages (notably English, Spanish, and Chinese) were rapidly acquiring new speakers, while half of the world's known languages faced extinction. Applications of linguistic anthropology seek remedies for language extinction and language-based discrimination, which are often driven by popular ideologies about the relative prestige and utility of different languages. (Ja.Hi.)

PSYCHOLOGICAL ANTHROPOLOGY

Psychological anthropology focuses on the mind, body, and subjectivity of the individual in whose life and experience culture and society are actualized. Within this broad scope there is no unified theoretical or methodological consensus, but rather lively debates about the relative importance of culture versus individual psychology in shaping human action and about the universality versus the inherent variability of human beings. The field unites a number of disparate research traditions with different intellectual programs, but it also provides an arena for principled argumentation about the existence of a common human nature.

Because of its focus on the individual who lives and embodies culture, psychological anthropological writing is often the study of one or a few actual people. Such "person-centred" ethnography augments a schematic view of cultural and social systems with a description and evocation of the experience of participating in such a system.

Researchers in the classical "culture-and-personality" school of psychological anthropology look for typical child-rearing customs, situations, patterns, or traumas that might result in characteristic responses (fantasies, anxieties, or conflicts) that in turn would find expression or resolution in the rituals, myths, and other features of the culture under study. Many employ a cross-cultural comparative methodology, seeking significant correlation between a childhood experience and adult institutions; for example, they look for a correlation between father-absence and harsh male initiation rites, thought necessary to counteract the strong maternal identification.

Ethnopsychiatry

Ethnopsychiatry examines not only other cultures' understandings of mental illness or abnormal states but methods of treatment other than standard Western procedures. Such systems as shamanism or spirit possession and the altered states of consciousness that accompany them are understood by some in terms of dissociation or schizoid states. For others these phenomena, often considered pathological in the West, are treated as normal in cultures that make productive use of human capabilities (or potentials) excluded from Western "folk psychology." (Rob.A.P.)

Sociology

Sociology studies human societies, their interactions, and the processes that preserve and change them. It does this by examining the dynamics of constituent parts of societies such as institutions, communities, populations, and gender, racial, or age groups. Sociology also studies social status or stratification, social movements, and social change, as well as societal disorder in the form of crime, deviance, and revolution.

Social life overwhelmingly regulates the behaviour of humans, largely because humans lack the instincts that guide most animal behaviour. Humans therefore depend on social institutions and organizations to inform their decisions and actions. Given the important role organizations play in influencing human action, it is sociology's task to discover how organizations affect the behaviour of persons, how they are established, how organizations interact with one another, how they decay, and, ultimately, how they disappear. Among the most basic organizational structures are economic, religious, educational, and political institutions, as well as more specialized institutions such as the family, the community, the military, peer groups, clubs, and volunteer associations.

Sociology, as a generalizing social science, is surpassed in its breadth only by anthropology—a discipline that encompasses archaeology, physical anthropology, and linguistics. The broad nature of sociological inquiry causes it to overlap with other social sciences such as economics, political science, psychology, geography, education, and law. Sociology's distinguishing feature is its practice of drawing on a larger societal context to explain social phenomena.

Explanation of social phenomena

Sociologists also utilize some aspects of these other fields. Psychology and sociology, for instance, share an interest in the subfield of social psychology, although psychologists traditionally focus on individuals and their mental mechanisms. Sociology devotes most of its attention to the collective aspects of human behaviour, because sociologists place greater emphasis on the ways external groups influence the behaviour of individuals.

The field of social anthropology has been historically quite close to sociology. Until about the first quarter of the 20th century, the two subjects were usually combined in one department (especially in Britain), differentiated mainly by anthropology's emphasis on the sociology of preliterate peoples. Recently, however, even this distinction has faded, as social anthropologists have turned their interests toward the study of modern culture.

Two other social sciences, political science and economics, developed largely from the practical interests of nations. Increasingly, both fields have recognized the utility of sociological concepts and methods. A comparable synergy has also developed with respect to law, education, and religion and even in such contrasting fields as engineering and architecture. All of these fields can benefit from the study of institutions and social interaction.

HISTORICAL DEVELOPMENT OF SOCIOLOGY

Though sociology draws on the Western tradition of rational inquiry established by the ancient Greeks, it is specifically the offspring of 18th- and 19th-century philosophy and has been viewed, along with economics and political science, as a reaction against speculative philosophy and folklore. Consequently, sociology separated from moral philosophy to become a specialized discipline. While he is not credited with the founding of the discipline of sociology, French philosopher Auguste Comte is recognized for having coined the term sociology.

The founders of sociology spent decades searching for the proper direction of the new discipline. They tried several highly divergent pathways, some driven by methods and contents borrowed from other sciences, others invented by the scholars themselves. A better view of the various turns the discipline has taken may be afforded by dividing the de-

velopment of sociology into four periods: the establishment of the discipline from the late 19th century until World War I, interwar consolidation, explosive growth from 1945 to 1975, and the subsequent period of segmentation.

Founding the discipline. Some of the earliest sociologists developed an approach based on Darwinian evolutionary theory. In their attempts to establish a scientifically based academic discipline, a line of creative thinkers, including Herbert Spencer, Benjamin Kidd, Lewis H. Morgan, E.B. Tylor, and L.T. Hobhouse, developed analogies between human society and the biological organism. They introduced into sociological theory such biological concepts as variance, natural selection, and inheritance—asserting that these evolutionary factors resulted in the progress of societies from stages of savagery and barbarism to civilization by virtue of the survival of the fittest. Some writers believed that these stages of society could be seen in the developmental stages of each individual. Strange customs were explained by assuming that they were throwbacks to useful practices of an earlier period, such as the make-believe struggle sometimes enacted between the bridegroom and the bride's relatives reflecting the earlier custom of bride capture.

Diminished interest in social Darwinism

In its popular period of the late 19th and early 20th centuries, social Darwinism, along with the doctrines of Adam Smith and Thomas Malthus, touted unrestricted competition and laissez-faire so that the "fittest" would survive and civilization would continue to advance. Although the popularity of social Darwinism waned in the 20th century, the ideas on competition and analogies from biological ecology were appropriated by the Chicago School of sociology (a University of Chicago program focusing on urban studies, founded by Albion Small in 1892) to form the theory of human ecology that endures as a viable study approach.

Replacing Darwinist determinism. Since the initial interest in evolutionary theory, sociologists have considered three deterministic theories to replace social Darwinism. This search for new approaches began prior to World War I as emphasis shifted from economic theory to geographic and cultural theory—roughly in that order.

Economic determinism. The first theory, economic determinism, reflects the interest many sociologists had in the thought of Karl Marx, such as the idea that social differentiation and class conflict resulted from economic factors. This approach had its greatest popularity in Europe, where it remained a strong influence on some sociologists until the 1980s. It did not gain a significant foothold in the United States, because American society was thought to be socially mobile, classless, and oriented to the individual. This neglect of Marxism by American sociologists, however, was not due to scholarly ignorance. Sociologists of all periods had read Marx as well as Charles A. Beard's economic interpretation of American history and the work of Werner Sombart (who had been a Marxist in his early career). Instead, in the 1960s, neo-Marxism—an amalgam of theories of stratification by Marx and Max Weber—gained strong support among a minority of sociologists. Their enthusiasm lasted about 30 years, ebbing with the breakup of the Soviet system and the introduction of postindustrial doctrines that linked class systems to a bygone industrial era. The persistence of social and economic inequality is now explained as a complex outcome of a number of factors, including gender, race, and region, as well as global trade and national politics.

Human ecology. Representing the second theoretical area, human geographers—Ellsworth Huntington, Ellen Semple, Friedrich Ratzel, Paul Vidal de La Blache, Jean Brunhes, and others—emphasized the impact of climate and geography on the evolution of those societies that flourished in temperate zones. Their theories found no place in mainstream sociological thought, however, except for a brief period in the 1930s when human ecology sought to explain social change by linking environmental conditions with demographic, organizational, and technological factors. Human ecology remains a small but vital part of sociology today.

Cultural theory. Finally, cultural theories of the 1930s emphasized human ability to innovate, accumulate, and diffuse culture. Heavily influenced by social and cultural anthropology, many sociologists concluded that culture was the most important factor in accounting for its own evolution and that of society. By 1940 cultural and social explanations of societal growth and change were accepted, with economic and geographic factors playing subsidiary roles.

EARLY SCHOOLS OF THOUGHT

Early functionalism. Scholars who established sociology as a legitimate social science were careful to distinguish it from biology and psychology, fields that had also begun to generalize about human behaviour. They did this by developing specific methods for the study of society. French sociologist Émile Durkheim (1858–1917), prominent in this regard, argued that various kinds of interactions between individuals bring about certain new properties (*sui generis*) not found in separate individuals. Durkheim insisted that these "social facts," as he called them—collective sentiments, customs, institutions, nations—should be studied and explained on a distinctly societal level (rather than on an individual level). To Durkheim the interrelations between the parts of society contributed to social unity—an integrated system with life characteristics of its own, exterior to individuals yet driving their behaviour. By positing a causal direction of social influence (from group to individual rather than the reverse, the model accepted by most biologists and psychologists of the time), Durkheim gave a much-needed framework to the new science of sociology. Some writers called this view "functionalism," although the term later acquired broader meanings.

Mechanical solidarity and organic solidarity

Durkheim pointed out that groups can be held together on two contrasting bases: mechanical solidarity, a sentimental attraction of social units or groups that perform the same or similar functions, such as preindustrial self-sufficient farmers; or organic solidarity, an interdependence based on differentiated functions and specialization as seen in a factory, the military, government, or other complex organizations. Other theorists of Durkheim's period, notably Henry Maine and Ferdinand Tönnies, made similar distinctions—status and contract (Maine) and Gemeinschaft und Gesellschaft (Tönnies)—and predicted that civilization would progress along the lines of specialization, contractual relations, and Gesellschaft.

Later anthropologists, especially Bronisław Malinowski and A.R. Radcliffe-Brown, developed a doctrine of functionalism that emphasized the interrelatedness of all parts of society. They theorized that a change in any single element would produce a general disturbance in the whole society. This doctrine eventually gained such a following among social anthropologists that some advocated a policy of complete noninterference, even with objectionable practices in preliterate societies (such as cannibalism or head-hunting), for fear that eliminating the practice might produce far-reaching social disorganization.

The functionalist-conflict debate. American sociology began undergoing significant development in the 1940s. The monumental growth of university enrollment and research after World War II was fueled by generous federal and private funding of research. Sociologists sought to enhance their status as scientists by pursuing empirical research and by conducting qualitative analysis of significant social problems. The struggle over the meaningful use of statistics and theory in research began at this time and remained a continuing debate in the discipline.

The gap between empirical research and theory persisted, in part because functionalist theory seemed divorced from the empirical research programs that defined mid-20th-century sociology. Functionalism underwent some modification when sociologist Talcott Parsons enunciated the "functional prerequisites" that any social system must meet in order to survive: developing routinized interpersonal arrangements (structures), defining relations to the external environment, fixing boundaries, and recruiting and controlling members. Along with Robert K. Merton and others, Parsons classified such structures on the basis of their functions. This approach, called structural-functional analysis (and also known as systems theory), was applied so broadly that Marion Levy and Kingsley Davis suggested it was synonymous with the scientific study of social organization.

That structural-functional emphasis changed in the 1960s, however, with new challenges to the functionalist notion that a society's survival depended on institutional practices. This belief, along with the notion that the stratification system selected the most talented and meritorious individuals to meet society's needs, was seen by some as a conservative ideology that legitimated the status quo and thereby prevented social reform. It also ignored the potential of the individual within society. In a response to the criticism of structural-functionalism, some sociologists proposed a "conflict sociology." In this view, the dominant institutions repress the weaker groups. This view gained prominence in the United States with the social turmoil of the civil rights struggle and the Vietnam War over the 1960s and '70s and prompted many younger sociologists to adopt this neo-Marxist view. Their interpretation of class conflict seemed consistent with the principal tenet of general conflict theory: that conflict pervades all of society, including the family, the economy, polity, and education.

Rising segmentation of the discipline. By 1975 the era of growth, optimism, and surface consensus in sociology had come to an end. The functionalist-conflict debate signaled further and permanent divisions in the discipline, and virtually all textbooks presented it as the main theoretical divide, despite Lewis A. Coser's widely known proposition that social conflict, while divisive, also has an integrating and stabilizing effect on society. Conflict is not necessarily negative, argued Coser in *The Functions of Social Conflict* (1936), because it can ultimately foster social cohesiveness by identifying social problems to be overcome.

Major modern developments. One of the consequences of the functionalist-conflict divide, recognized by the 1970s as unbridgeable, was a decline in general theory building. Others were growing specialization and controversy over methodology and approach. Communication between the specialties also diminished, even as ideological disputes and other disagreements persisted within the specialty areas. New academic journals were introduced to meet the needs of the emerging specializations, but this further obscured the core of the discipline by causing scholars to focus on microsociological issues.

Social stratification. Since social stratification is the most binding and central concern of sociology, changes in the study of social stratification reflect trends in the entire discipline. The founders of sociology—including Weber—thought that the United States, unlike Europe, was a classless society with a high degree of upward mobility. During the Great Depression, however, Robert and Helen Lynd, in their famous Middletown (1937) studies, documented the deep divide between the working and the business classes in all areas of community life. Likewise, C. Wright Mills in 1956 proposed that a "power elite" dominated the national agenda in Washington, a cabal comprising business, government, and the military.

General theory and systems theory

Attempting to build a general theory, Gerhard Lenski shifted attention to whole societies and proposed an evolutionary theory in *Power and Privilege* (1966), demonstrating that the dominant forms of production (hunting and gathering, horticulture, agriculture, and industry) were consistently associated with particular systems of stratification. Addressing the contemporary world, Marion Levy theorized in *Modernization and the Structures of Societies* (1960) that underdeveloped nations would inevitably develop institutions that paralleled those of the more economically advanced nations, which ultimately would lead to a global convergence of societies. Challenging the theory as a conservative defense of the West, Immanuel Wallerstein's *The Modern World System* (1974) proposed that advanced industrial nations would develop most rapidly and thereby widen global inequality by holding the developing nations in a permanent state of dependency.

Having been challenged as a male-dominated approach, traditional stratification theory was massively reconstructed in the 1970s to address the institutional gender inequalities found in all societies. Rae Lesser Blumberg, drawing on the work of Lenski and economist Esther Boserup, theorized the basis of persistent inequality in *Stratification, Socioeconomic, and Sexual Inequality* (1978). Janet Saltzman Chafetz took economic, psychological, and sociological factors into account in *Gender Equity: An Integrated Theory of Stability and Change* (1990). Traditional theories of racial inequality were challenged and revised by William Julius Wilson in *The Truly Disadvantaged* (1987).

Disciplinary specialization, especially in the areas of gender, race, and Marxism, came to dominate sociological inquiry. Sociologists have increasingly turned to large-scale surveys and government data banks as sources for their research. Social stratification theory and research continue to undergo change and have seen substantive reappraisal ever since the breakup of the Soviet system.

Interdisciplinary influences. The significant growth of sociological inquiry after World War II prompted interest in historical and political sociology. Charles Tilly in *From Mobilization to Revolution* (1978), Jack Goldstone in *Revolutions: Theoretical, Comparative and Historical Studies* (1993), and Arthur Stinchcombe in *Constructing Social Theories* (1987) made comparative studies of revolutions and proposed structural theories to explain the origins and spread of revolution.

Improved study of collective action

From its inception the study of social movements looked closely at interpersonal relations formed in the mobilization phase of collective action. Beginning in the 1970s, scholars focused more deeply on the long-term consequences of social movements, especially on evaluating the ways such movements have propelled societal change. The rich area of historical and international research that resulted includes the study of social turmoil's influence on New Deal legislation; the rise, decline, and resurrection of women's rights movements; analysis of both failed and successful revolutions; the impact of government and other institutions on social movements; national differences in how social movements spur discontent; the response of nascent movements to political changes; and variations in the rates of growth and decline of movements over time and in different nations. In short, countering the general trend, social movement research became better integrated into other specialties, especially in political and organizational sociology.

Stratification studies and organizational sociology were broadened to include economic phenomena such as labour markets and the behaviour of businesses. Econometric methods were also introduced from economics. Thus, to predict income, sociologists would examine status variables (such as race, ethnicity, or gender) or group affiliations (looking at degree of unionization, whether groups are licensed or unlicensed, and the type of industry, community, or firm involved). Other economic variables tapped by sociologists include human capital (education, training, and experience) and economic segmentation. As a result of his interaction with economists, for example, James S. Coleman was the first sociologist since Parsons to build a comprehensive social theory. Coleman's *Foundations of Social Theory* (1990), based on economic models, suggests that the individual makes rational choices in all phases of social life.

The historical divide: qualitative and establishment sociology. Paradoxically, American sociology, unlike its European counterpart, has been marked by an individualistic (psychological) orientation, even though early sociologists fought to establish a discipline distinct from psychology. Most specialized research in American sociology still uses the individual as the unit of analysis. The standard practice is to collect data from or about individuals, categorize their social characteristics into "groups," and relate them to other categories of individuals such as income classes, occupations, and age groups. These intergroup relations are often examined with complex statistical tools. This practice is not generally recognized as social-psychological in nature, yet neither is it regarded as social structural analysis. Only a minority of sociologists in fields such as demography, human ecology, and historical or comparative institutional study use actual groups, organizations, and social structures as units of analysis.

As the field developed in the United States, many early 20th-century sociologists rejected instinctivist psychology and the classical behaviourism of John B. Watson. One group, however, emphasized the study of individuals in an

approach called symbolic interaction, which took root at the University of Chicago early in the 20th century and remains prominent in contemporary sociology. John Dewey, George H. Mead, and Charles H. Cooley argued that the self is the individual's internalization of the wider society as revealed through interaction, the accumulated perceptions of how others see them. In other words, the mind and human self are not innate human equipment but constructions of the "person" (the socialized individual) derived from experience and intimate interpersonal interaction in small groups. This constructed self, however changing, functions as a guide to social behaviour. Social reality is thus made up of constructed symbols and meanings that are exchanged with others through daily interaction.

Symbolic interaction

William I. Thomas and Ellsworth Faris used symbolic interaction theory to guide their empirical research in the tradition of Robert E. Park and Ernest W. Burgess by using personal documents, life histories, and autobiographies. The two revealed how people attach meanings to their experience and to the broader social world. This research tradition was enriched after 1960 by several innovations. The most sophisticated small-group research was devised by Erving Goffman in *The Presentation of Self in Everyday Life* (1959). Goffman insisted that the most meaningful individual behaviour occurs in the chance, intimate encounters of each day. These encounters include greeting people, appearing in public, and reacting to the physical appearance of others. Such encounters have structures of their own that can be researched by carefully constructing the "frames" (points of reference) people use to interpret and "stage" interactions. The structures are thought to represent true reality as opposed to the artificially constructed concepts that sociologists impose on the subjects they study.

In *Studies in Ethnomethodology* (1967), Harold Garfinkel coined the term "ethnomethodology" to designate the methods individuals use in daily life to construct their reality, primarily through intimate exchanges of meanings in conversation. These constructions are available through new methods of conversational analysis, detailed or "thick" descriptions of behaviour, "interpretive frames," and other devices. Proponents of this view have favoured the work of earlier European phenomenology, Verstehen (historical understanding), and interpretive sociology. More recently, qualitative sociologists have drawn on French structuralism, poststructuralism, and postmodernism to emphasize ways the "deeper" sources of hidden meanings in culture and language can affect the behaviour of individuals or of whole societies.

Thick description

Since World War II, sociology has exported much of its theory, methodology, and findings to other divisions of the university, sometimes to its disadvantage. The study of human relations and formal organizations was transferred to business schools. The study of socialization, institutions, and stratification was absorbed by departments of education. Outside the university, the empirical methods and sociological theory prompted government agencies to adopt a behavioral perspective. Economists widened the scope of their research by introducing social variables to the analysis of economic behaviour. In short, although contemporary sociology is divided, it remains a vibrant field whose innovations contribute to its own development and that of social science in general.

METHODOLOGICAL CONSIDERATIONS IN SOCIOLOGY

Much 19th-century sociology had no system for gathering and analyzing data, but over time the inadequacies of speculative methods became increasingly evident, as did the need for obtaining reliable and verifiable knowledge. Like his contemporaries, Herbert Spencer assembled vast stores of observations made by others and used these to illustrate and support generalizations he had already formulated. Early social surveys like those conducted by Charles Booth in a monumental series on the social problems of London produced masses of data without regard to their theoretical relevance or reliability. Frédéric Le Play made similar use of the French case studies he drew on for his extensive investigations of family budgets.

Early exploitation of statistical materials, such as official records of birth, death, crime, and suicide, provided only moderate advances in knowledge. Data were easily manipulated, often to support preconceived ideas (the status quo). Among the most successful of such studies was that on suicide rates by Durkheim in *Le Suicide* (1897). Moreover, his *Rules of Sociological Method* (1895) had begun to meet the standards of scientific inquiry. In gathering data on suicides, Durkheim considered the social characteristics of individuals (*e.g.*, religious affiliation, rural-urban residence) that reflected the degree of their social integration in the community, and he related these variables statistically.

Methodological development in contemporary sociology. At the beginning of the 20th century, interest in developing a sociological methodology grew steadily. Methodological approaches outlined in W.I. Thomas and Florian Znaniecki's *Polish Peasant in Europe and America* (vol. 5, 1918–20) were recognized as an important advance, not so much in methodology as in committing sociologists to the task of improving methodology. Thomas and Znaniecki systematically gathered longitudinal data through letters, diaries, life histories, and other relevant documents. Intended to gather specific data to help planners solve social problems, this approach soon became popular. The most ambitious of these "community social surveys" was the two-volume work *Great Depression, Recent Social Trends* (1933), edited by sociologists W.F. Ogburn and H.W. Odum.

Significant advances in scientific methodology occurred at the University of Chicago in the 1920s. Many studies of the metropolis and its subareas were conducted under the leadership of Robert E. Park, Ernest Burgess, and their colleagues. Most important, hypotheses were developed during the research rather than being imposed a priori (a practice later replaced by theoretically guided research). Many students took part in the studies and contributed to methods and findings.

Ecological patterning. A critical aspect of the Chicago School's urban research involved mapping locations. These included locations of land values, specific populations (racial, ethnic, or occupational), ethnic succession in neighbourhoods, residences of persons who committed certain crimes, or zones with a high incidence of divorce and desertion. Data collection methods included participant observation, life histories, case studies, historical information, and life cycles of social movements. Sociologists at the University of Chicago paid equal attention to the improvement of methodology as they developed this approach. Here, for the first time, was a large-scale effort in which theory, methodology, and findings evolved together in an inductive process. Growing from its success in the American Midwest, urban research and zone mapping spread throughout the United States and influenced sociology abroad.

Urban studies

Ecological methods such as urban mapping were also first developed at Chicago, having grown out of the research on the metropolitan region, especially that regarding nonsocial patterns that resulted from the movement of populations, businesses, industries, residences, and institutions as each sought more advantageous locations. Most early urban studies mapped distributions that revealed relationships in general patterns of urban ecology. Because of this, multiple indicators of disorganization, stratification, vertical mobility, and population phenomena were found to follow regularities and could be considered predictable to some degree (see DEMOGRAPHY in the *Micropædia*).

Experiments. Experimental methods, once limited to the domain of psychologists and considered inapplicable to social research, were eventually applied to the study of groups. By the 1930s, social psychologists Kurt Lewin and Muzafer Sherif and their colleagues began conducting experiments on social interaction. Sociologists soon followed their example and set up research laboratories. Notably, Robert F. Bales at Harvard systematically observed interaction in small artificial groups, producing useful results that were replicated elsewhere.

As a rule, successful experiments tend to occur in simple situations in which the variables are limited or controlled. Complex experiments, however, are possible. At Stanford, for example, a series of experiments over 30 years con-

tributed to a formal theory of social status building and maintenance set forth by Joseph Berger and Morris Zelditch in *Status, Rewards, and Influence* (1985). At the University of Iowa, two decades of laboratory and computer-simulated research on power and exchange in small groups advanced theory in networks and decision making, summarized by Barry Markovsky in *Social Psychology of Small Groups* (1993).

Statistics and mathematical analysis. Sociologists have increasingly borrowed statistical methods from other disciplines. Statistician Karl Pearson's "coefficient of correlation," for example, introduced an important concept for measuring associations between continuous variables without necessarily defining the nature of the connection. Later, statistical estimates of causal relations were probed by "multiple regression analysis," employing techniques that estimate the degree to which any particular variable influences a particular outcome.

Patterns of responses to interview questions, once thought to be purely qualitative, have also been subject to mathematical scaling. A method devised by psychologist L.L. Thurstone in the late 1920s gained popular use in sociology. In this approach a list of items is presented to a number of judges who individually relist them in order of importance or of interest. Items on which there is substantial agreement are then reordered to form a scale. Another technique asks participants to respond to statements by strength of agreement (strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree). Social distance may be measured by asking respondents whether they would accept members of other groups as spouses, close friends, fellow employees, neighbours, or citizens.

A method called sociometry, introduced by J.L. Moreno in the 1930s, collects and tabulates information on group interactions. The interactions studied can appear trivial—for example, who confides in whom, which friends eat lunch together—or they may be more businesslike, such as who might be appointed as a group spokesperson. The data may be gathered by direct observation, interviews, or questionnaires. The preferences each individual has for specific others are then mapped with arrows from sender to receiver, and this results in a diagram of choices for the entire group. The person chosen most often is labeled a "star" or, less often, an "isolate."

Mapping social structures

The patterns may be quantified and supplemented with other data to reveal a group's informal structure. A powerful application of the approach, often mathematized, called network analysis, maps different types of interactions between organizations over extended periods and thus exposes a substructure not revealed from organizational charts or public documents.

Factor analysis, an elaboration of Pearson's coefficient of correlation, significantly reduces the number of complex variables to be considered. For instance, 50 different questions or measures of work alienation may in fact represent only seven or eight dimensions of alienation. Factor analysis reduces the variables to a more practical number of common factors and then determines each factor's relative contribution to the outcome variables.

Many other statistical methods have been devised to suit the purposes of such specialties as demography, ecology, social stratification, organizational analysis, mass communication, and social movements. Statistical methods have developed so rapidly that they sometimes outstrip scholars' ability to find data worthy of their application. Computers have accelerated the application of complex measures that were previously limited by the amount of time required for performing the mathematics. Progress in measurement has been so significant that the American Sociological Association in 1969 established an annual volume entitled *Sociological Methodology*.

Data collection. Research techniques vary depending on the social phenomena studied. Data can be obtained through participant observation, content analysis, interviewing, and documentary analysis. Each study requires a specific unit of observation, be it an individual, an organization, a city, a relationship between units, or a statistical rate. Even the way a concept is defined can affect data collection.

Data collection

Steps must be taken to collect valid data. Many obstacles can arise, especially on sensitive subjects such as alcohol consumption in a community that prohibits or looks down upon it. In this instance the problem of gathering valid data might be circumvented by counting liquor bottles in trash receptacles or in the town dump. Similarly, a decline in the number of fictional works checked out of libraries has been used to estimate television-watching habits. Unfortunately, questionnaires, while useful for gathering information from large numbers of respondents, are marked by methodological problems. The wording of questions must be intelligible to the uneducated or uninterested as well as to the sophisticated respondent. Topics that provoke resistance must be presented in a way that yields a complete and unbiased response while keeping the interviewee engaged with the questions.

In face-to-face interviewing, it may be necessary to consider the interviewer's sex or race, appearance, manner, and approach. Questions must be posed in a way that does not influence the response. Interviewers must have steps for handling resistance or refusal. Indirect questioning, for example, may yield information that respondents would hesitate to provide in answers to direct questioning. Because of this, information collected through "canned" telephone interviews often leads to lower-quality data and poorer response rates.

Sampling errors and bias both constitute a continuing concern, especially since so much sociological knowledge is derived from samples of a larger universe. Where bias cannot be controlled, its extent may sometimes be estimated by various methods, including intensive analysis of smaller samples.

National methodological preferences. Research approaches vary from country to country. In France, Italy, and several other European nations, industrial sociology is understandably important, much of it based on case studies of industries and the experiences of workers. Sociology in Britain, the Scandinavian countries, and Japan covers most of the fields mentioned above. For most western European countries, interest focuses on social stratification and its political implications.

In fact, general differences between the sociologies of European countries and that of the United States were established early in the 20th century. The European approach favoured broad sociological theory based on philosophical methods, while the American approach favoured induction and empiricism. Sociology in Russia and eastern European countries is also becoming more similar to its Western counterparts. Research in the former Soviet-bloc nations, previously shaped by the concepts and methods of Marxist sociology, has shifted to approaches influenced by European and American sociology.

Regional and ideological differences

More important than national preference is the methodological divide between scientific sociology and applied sociology; scholars interested in applied sociology tend to deprecate the methods and findings of the scientific sociologists as either irrelevant or ideologically biased. Issues of ethics have also been raised, particularly regarding observations and experiments in which the privacy of subjects may be felt to be invaded.

Finally, the divide between mainstream sociologists and those devoted to qualitative analysis seems deep and unbridgeable. Qualitative sociologists feel that their work is underrecognized and marginalized, even though it deals more with social reality than does standard sociology. Classical sociologists, in turn, feel that qualitative work can be trivial, philosophical, ideologically driven, or difficult to research. In addition, some members of groups who feel exploited (women, blacks, homosexuals, and the working class) assert that social observations cannot be made by outsiders; they believe that only victims have true insight into other victims and that they alone are equipped to do meaningful research in these areas. Minorities and other groups that locate themselves at the margins of society sometimes come together—often by organizing movements within professional societies—to challenge "establishment sociologists." This results in the direction of more attention, funding, and research to the more highly focused topics.

STATUS OF CONTEMPORARY SOCIOLOGY

Academic status. The Greek philosophers and their European successors discussed much of the subject matter of sociology without thinking of it as a distinct discipline. In the early 19th century, the subject matter of the social sciences was discussed under the heading of moral philosophy. Even after Comte introduced the word sociology in 1838, sociological studies were combined with other subjects for some 60 years. Not until universities undertook a commitment to the subject could one person make a living as a full-time sociologist. This commitment had to be made first by scholars in other fields such as history and economics.

United States. As early as 1876, at the newly established Johns Hopkins University, some sociology was taught in the department of history and politics. In 1889 at the University of Kansas, the word appeared in the title of the department of history and sociology. In 1890 at Colby College, historian Albion Small taught a course called sociology, as did Franklin H. Giddings in the same year at Bryn Mawr College. But the first real commitment to the creation of a field of sociology took place in 1892 at the then new University of Chicago, where the recently arrived Small received permission to create a department of sociology—the first such in the world. Within two years sociology departments had been founded at Columbia, Kansas, and Michigan, and shortly thereafter they were begun at Yale, Brown, and many other universities. By the late 1890s nearly all higher educational institutions in the United States either had departments of sociology or offered courses in the subject.

In 1895 the *American Journal of Sociology* began publication at the University of Chicago; in time a large number of journals followed in many other countries. Ten years later the American Sociological Society was organized, also to be followed by a large number of national, regional, international, and specialized sociological organizations. These groups institutionalized the subject and continue to guide its directions and define its boundaries. Eventually in 1949 the International Sociological Association was established under the sponsorship of UNESCO, and Louis Wirth of the University of Chicago was elected its first president. The rapid increase of full-time sociologists, along with the growth of sociology publications, allowed the content of the discipline also to expand rapidly. Research grew throughout the 20th century at an accelerated pace, especially after World War II, partly because of strong financial support from foundations, government, commercial sources, and individuals. This period was also marked by the rising popularity of anthropology, and many universities formed joint anthropology-sociology departments. By the 1960s, however, growing interest in anthropology had resulted in the formation of separate anthropology departments at the larger research universities. At the same time, interest in sociological research continued to develop. By 1970 there were more than a dozen important sociological journals and an indefinite number of minor journals worldwide. Along with this growth came a flourishing of research institutions—some affiliated with university departments and some independent—which allowed a small but increasing number of sociologists to pursue full-time research free from teaching responsibilities.

Developments in sociological research

France. In France, where Comte and later Durkheim gave early impetus to sociology, sociological research developed in a number of fields. The two world wars slowed that development somewhat, but after 1945 a strong revival of interest in sociology took place, during which the French government established a number of research institutes in the social sciences parallel to those in the natural sciences, including several in Paris—notably the Centre d'Études Sociologiques, the Institut National d'Études Démographiques, and the Maison des Sciences de l'Homme. These government-funded institutes employ many full-time sociologists, some of them among the more prominent scholars in the nation. The growth of sociological research at French universities has been somewhat more conservative; the Sorbonne, for example, in 1970 had only one chair officially assigned to sociology. The University of Nanterre, however, established a department with four professorships.

Germany. German sociology had a strong base in the late 19th century and afterward, and the writings of Tönnies, Weber, Georg Simmel, and others had an international impact. By the early 1930s, however, official Nazi hostility had impeded German sociology's development, and by the time of World War II the Nazis had destroyed sociology as an academic subject. Immediately after the war a new generation of scholars, aided by visiting sociologists, imported the new empirical research methods and began to develop a style of German sociology much different from the earlier theoretical and philosophical traditions. At the University of Frankfurt, Max Horkheimer's Institut für Sozialforschung (social research), established by private financing before the war, was revived. The University of Cologne also established a department notable for its survey research. West German universities remained conservative for a time, but two newly created ones—the Free University of Berlin and the University of Constance—made sociology one of their major disciplines. By 1970 most West German universities had at least one chair in sociology. National needs received special emphasis, including studies of unemployment, youth problems, and delinquency. A significant amount of German research also is published in such fields as rural sociology, political sociology, and the family.

United Kingdom and Commonwealth. Despite the early prominence of Herbert Spencer and L.T. Hobhouse, the leading universities of the United Kingdom virtually ignored sociology until the mid-20th century. Before World War II, Britain excelled in anthropology, especially in the study of the British Empire's nonwhite societies. British sociology concentrated on studies of the poor, and much of it was undertaken by people with experience in social work rather than social research. The major prewar sociology department, at the London School of Economics, prioritized social reform over scientific research. In the postwar period, however, a considerable revival of sociology took place; Oxford and Cambridge recognized the subject by creating positions for sociologists, and various new universities established chairs and departments. Significant work in Britain has emerged in such fields as population and demography, sociology of organization, politics and industry, social stratification, and general sociology. The Tavistock Institute of Human Relations in London has become world famous and concentrates on human relations in the family, the work group, and organizations.

A parallel growth took place in Canada, Australia, and New Zealand. Canada, with some apparent reluctance, allowed itself to be much influenced by American sociology and has built many new departments with sociologists trained in the United States.

Scandinavia and The Netherlands. To a considerable extent Scandinavia and The Netherlands have also adopted the methods and some of the content of American sociology, and the subject has developed rapidly at universities and research institutes. There is also a considerable exchange between sociologists in these countries, because their works are typically published regionally as well as in the United Kingdom, the United States, and Germany.

Japan. Japanese interest in sociology dates back to the 1870s. The Japanese Sociological Society (Nippon Shakai Gakkai), headquartered at the University of Tokyo, was founded in 1923; by 1960 there were about 150 universities and colleges with courses in the subject. In the early period sociology was nearly all imported; Comte and Spencer, and later Giddings and Gabriel Tarde, were the most influential theorists. After World War II there were rapid changes in sociology in Japan, with empirical research methods largely replacing the earlier philosophical approach. Importations from American sociology were abundant. Popular among these were industrial sociology, social stratification, educational sociology, public opinion research, and the study of mass communication.

Emphasis on empirical methods

Soviet bloc. Sociology in the former Soviet Union was long held back by the perceived incompatibility of the subject with Marxist theory. Eventually it was permitted to develop, and the number of sociological institutes and chairs of sociology increased. By 1970 the Soviet Sociological Association had more than a thousand members. Leading research interests included labour productivity, education, crime, and alcoholism. Soviet sociology generally avoided issues that might have implied conflict with Marxist

thought, concentrating for a time on demography and time-budget studies.

The nations of the Soviet bloc were also periodically inhospitable to sociology, but the strong interest of younger scholars alleviated some of this opposition, and in the second half of the 20th century sociology made considerable progress in Hungary, Poland, the Czech Republic, and Slovakia.

Israel. In Israel the dominant department of sociology is at the Hebrew University in Jerusalem, where there are also several research institutes. Departments were also established at the University of Haifa and Tel Aviv University. Israeli sociology maintains continuous close contacts with American sociology, and many of the leading Israeli sociologists have trained or taught in the United States. Among the specialties of Israeli sociology are research in methodology, communication, and criminology. Similarly prominent is the study of collective settlements (kibbutzim), in which new forms of custom and social organization are observed as they develop. Studies of stratification and the labour market have also explored the inequality between Israelis and Arabs.

Italy. In Italy, interest in sociology developed in the mid-20th century at several universities, and academic chairs and research institutes gradually increased. Of particular interest to Italian sociology are studies of industrial efficiency, social movements, and social mobility. The model of centralized control over universities, however, has hindered the development of the discipline, both in Italy and in Spain.

Latin America. In Latin America objective sociology was long resisted, partly because it was viewed as a threat to the political and social order but also because of meagre financial support for research and the low salary level of professors, many of whom were forced to supplement their earnings by engaging in other occupations. In the 1960s, however, the number of full-time chairs increased, and a number of research institutes, some financed by U.S. funds, were established. Political instability in some countries remains a major hindrance, and in such countries able scholars continue to be forced from their university positions from time to time.

Less-developed nations. Little by little, sociology has penetrated some of the less-developed nations. A number of African universities have formed departments, and the subject is gaining in importance in the Philippines, India, Indonesia, and Pakistan. Some of the more significant developments have occurred in India, where a number of important research institutes have been established.

Comparisons with established disciplines

Scientific status. Sociology has not achieved triumphs comparable to those of the older and more heavily supported sciences. Several interpretations have been offered to explain the difference—most frequently, that the growth of sociological knowledge is more random than cumulative. Yet, in some parts of the discipline—such as methodology, human ecology, demography, social differentiation and mobility, attitude research, small-group interaction, public opinion, and mass communication—a slow but significant accumulation of organized and tested knowledge has taken hold. By comparison, some other fields lack this expanding volume of literature. Still, the slow development of published sociological research may stem from a variety of factors: excess use of jargon, a disposition for pseudoquantification, excessive imitation of natural science methodology, and overdependence on interview data, questionnaires, or informal observations. Contemporary sociology is indeed marked by all these shortcomings, but in general there has been progress toward clearer communication and improved methodology, both of which yield more reliable data. As a result, conclusions are drawn from research methods applied to replicated studies that are, in turn, less dependent on the strength of one particular methodological device.

Bias is sometimes presumed to be a chronic affliction of sociology. This may arise in part from the fact that the subject matter of sociology is familiar and important in everyone's daily life. As a result, variations in philosophical outlook and individual preferences can contribute to an irrational bias. Thus, critics have expressed disapproval of the sociologists' skepticism on various matters of faith, of their amoral relativism concerning customs, of their apparent oversimplifications of some principles, and of their particular fashions in categorization and abstraction. But skepticism toward much of the content of folk knowledge is a characteristic of all science, and relativism can be interpreted as merely an avoidance of antiscientific ethnocentrism. Furthermore, abstraction, categorization, and simplification are necessary to the advancement of knowledge, and no one system satisfies everyone.

The purpose of sociology

The dispute about the main purpose of sociology—whether it works to understand behaviour or to cause social change—is a dispute found in every pursuit of scientific knowledge, and such polarization is far from absolute. Scholars differ in the degree to which they regard the value of science as an intellectual understanding of the cosmos or as an instrument for immediate improvement of the human lot. Since even the "purest" scientist conceives of his work as benefiting mankind, the issue narrows to a difference in preference between an ad hoc attack on immediate human problems and a long-run trust that basic knowledge, gathered without reference to present urgencies, is even more valuable. In some countries there is much pressure toward early practicality of results; in others, including the United States, the larger number of scholars and the principal sociological associations have shown preference for "basic science."

A degree of polarization has also arisen over the proper strategy for research—whether research should take its direction from the needs of society and humankind or from the evolving theoretical corpus of sociology. In nations that allow academic freedom, such disputes are usually of low intensity, because scholars select research interests on any basis they prefer, including that of personal taste. In this way presumably the motivation of the investigator is maximized.

Sociologists most interested in action express impatience at the claims of others who prefer to separate their research from personal values. Much of the dispute prevails only because the two sides argue past each other. There can be wide agreement that no human being is without personal values, that research forced to confirm a particular set of values is not good science, and that there can be scientific issues toward which a particular investigator is value-neutral. In research that is susceptible to contamination by the values of the worker, it is generally possible to minimize the damage by employing methodological devices that prevent the researcher from imposing his or her wishes on a particular outcome. These devices include objective observational techniques, and measurement methods and independent or blind analysis of results.

Current trends. Sociology will continue to grow in the foreseeable future. Among present trends contributing to this growth are the increase in public appreciation of the subject, the continuing growth of funds for teaching and research, the steady reduction of sectarian opposition to study of social institutions, the refinement of methodologies that permit statistical analysis, and the growth of acceptance from scientists in other fields. Although factors such as extreme nationalism and internal conflict can inhibit growth in sociology, such conditions have impeded development only locally and temporarily.

Furthermore, it appears likely that public interest in the development of sociological knowledge will increase as more people come to realize what sociology can contribute to human safety and welfare. Advances in science and technology will always be accompanied by unforeseen and unintended consequences. Progress can indeed diminish the effects of natural catastrophes such as famine and disease, but progress can also bring about a wide range of new problems. These are not menaces of an impersonal nature but dangers that arise from imperfection in human behaviour, particularly in organized human relations. In addition, wars have shown a tendency to become larger and ever more destructive, and the causes, though far from being understood, clearly lie, in large measure, in the complexities of social organization and in the interaction of great corporate national bodies. It can be argued that politics, unaided by social science and other disciplines, cannot reverse this trend.

Increased understanding of social interaction

Problems within nations are seen as increasing sources of human troubles. There is a general rise in the severity of ethnic hostilities and of internal conflicts between generations, political factions, and other divisions of the populations. Human welfare is also threatened by widespread poverty, crime, vice, political corruption, and breakdowns in the family and in other institutions. Contemporary sociology does not yet provide the solutions, but its practitioners believe that the prospects for human betterment depend in large part on the increasing application of social science knowledge to these enduring problems.

Applications of sociology also appear to be spreading in several directions. Many sociologists are employed by national and international bodies to recommend programs, evaluate their progress and effects, gather data for planning, and propose methods for initiating change. Sociologists aid industry by obtaining data on clients and workers. Some of this work includes social surveys, offering advice on personnel or public relations problems, providing labour unions with advice, helping communities undertake reform, counseling families, and donating or selling advice to consumer groups. As long as organizations need information on their various publics, there will be strong demand for sociological knowledge.

Progress into the deeper sociological questions will require greater resources, larger research teams, and special research agencies. This compares to the increased complexity of research organization that occurred in the older sciences. In addition, large-scale sociological research will continue to be enhanced by the availability of computers, by complex statistical techniques, and by the storage capacity of data banks.

Emerging roles for sociologists. The principal employment of sociologists has been in academic institutions, but other employment possibilities have opened in recent decades. Social welfare agencies have long employed sociologists, and government organizations of all types—from bureaus dealing with population, budgets, and education to departments concentrating on crime, agriculture, and health matters—have tapped sociologists for help in research, planning, and administration. Other directions of sociological activity include the roles of consultant, social critic, social activist, and even revolutionary. When the activity diverges far enough from true scholarship and traditional academic sociology, it may cease to be regarded as sociological, but it appears likely that sociologists will continue to spread their activities over an ever-widening region of national or global concern. (R.E.L.F./Wi.F.)

RELATED FIELDS

Social psychology. Social psychology is the scientific study of the behaviour of individuals in their social and cultural setting. Although the term may be taken to include the social activity of laboratory animals or those in the wild, the emphasis here is on human social behaviour.

Once a relatively speculative, intuitive enterprise, social psychology has become an active form of empirical investigation, the volume of research literature having risen rapidly after about 1925. Social psychologists now have a substantial volume of data covering a range of topics; the evidence remains loosely coordinated, however, and the field is beset by many different theories and conceptual schemes.

Early impetus in research came from the United States, and much work in other countries has followed U.S. tradition, though independent research efforts are being made elsewhere in the world. Social psychology is being actively pursued in the United Kingdom, Canada, Australia, Germany, The Netherlands, France, Belgium, Scandinavia, Japan, and Russia. Most social psychologists are members of university departments of psychology; others are in departments of sociology or schools of business or work in such applied settings as industry and government.

Scope of social psychology

Much research in social psychology has consisted of laboratory experiments on social behaviour, but this approach has been criticized in recent years as being too stultifying, artificial, and unrealistic. Much of the conceptual background of research in social psychology derives from other fields of psychology. While learning theory and psychoanalysis were once most influential, cognitive and linguistic approaches to research have become more popular; sociological contributions also have been influential.

Social psychologists are employed, or used as consultants, in setting up the social organization of businesses and psychiatric communities; some work to reduce racial conflict, to design mass communications (*e.g.*, advertising), and to advise on child rearing. They have helped in the treatment of mental patients and in the rehabilitation of convicts. Fundamental research in social psychology has been brought to the attention of the public through popular media.

Research methods. Laboratory experiments, often using volunteer students as subjects, omit many features of daily social life. Such experiments also have been criticized as being subject to bias, since the experimenters themselves may influence the results. Research workers who are concerned more with realistic settings than with rigour tend to leave the laboratory to perform field studies, as do those who come from sociological traditions. Field research, however, also can be experimental, and the effectiveness of each approach may be enhanced by the use of the methods of the other.

Many colleges and universities have a social-psychology laboratory equipped with observation rooms permitting one-way vision of subjects. Sound and video recorders and other devices record ongoing social interaction; computing equipment and other paraphernalia may be employed for specific studies.

Social behaviour is understood to be the product of innate biological factors resulting from evolution and of cultural factors that have emerged in the course of history. Early writers (*e.g.*, William McDougall, a psychologist) emphasized instinctive roots of social behaviour. Later research and writing that tended to stress learning theory emphasized the influence of environmental factors in social behaviour. In the 1960s and '70s field studies of nonhuman primates (such as baboons) drew attention to a number of similarities to human social behaviour, while research in cultural anthropology has shown that many features of human social behaviour are the same regardless of the culture studied. It is coming to be a widely accepted view that human social behaviour seems to have a biological basis and to reflect the operation of evolution as in the case of patterns of emotional expression and other nonverbal communication, the structure of language, and aspects of group behaviour.

Much research has been done on socialization (the process of learning from a culture), and learning has been found to interact with innate factors. An innate capacity for language, for example, makes it possible to learn a local language. Culture consists of patterns of behaviour and ways of organizing experience; it develops over the course of history as leaders and innovators introduce new elements, only some of which are retained. Many aspects of social behaviour can be partly accounted for in terms of their history.

Social perception. In some laboratory experiments, subjects watch stills or moving pictures, listen to tape recordings, or directly observe or interact with another person. Subjects may be asked to reveal their social perception of such persons on rating scales, to give free descriptions of them, or to respond evaluatively in other ways. Although such studies can produce results that do not correspond to those in real-life settings, they can provide useful information on the perception of personality, social roles, emotions, and interpersonal attitudes or responses during ongoing social interaction.

The effect of cultural stereotypes

Research has been directed to how social perception is affected by cultural stereotypes (*e.g.*, racial prejudice), by inferences from different verbal and nonverbal cues, by the pattern of perceptual activity during social interaction, and by the general personality structure of the perceiver. The work has found practical application in the assessment of employees and of candidates for positions.

There also has been research on the ways in which perception of objects and people is affected by social factors such as culture and group membership. It has been shown, for example, how coins, colours, and other physical cues are categorized differently by people as a result of their

group membership and of the categories provided by language. Other studies have shown the effect of group pressures on perception.

Interaction processes. The different verbal and nonverbal signals used in conversation have been studied, and the functions of such factors as gaze, gesture, and tone of voice are analyzed in social-interaction studies. Social interaction is thus seen to consist of closely related sequences of nonverbal signals and verbal utterances. Gaze has been found to perform several important functions. Laboratory and field studies have examined helping behaviour, imitation, friendship formation, and social interaction in psychotherapy.

Among the theoretical models developed to describe the nature of social behaviour, the stimulus–response model (in which every social act is seen as a response to the preceding act of another individual) has been generally found helpful but incomplete. Linguistic models that view social behaviour as being governed by principles analogous to the rules of a game or specifically to the grammar of a language have also attracted adherents. Others see social behaviour as a kind of motor skill that is goal-directed and modified by feedback (or learning), while other models have been based on the theory of games, which emphasizes the pursuit and exchange of rewards and has led to experiments based on laboratory games.

Small social groups. All small social groups do not function according to the same principles, and, indeed, modes of social activity vary for particular kinds of groups; *e.g.*, for families, groups of friends, work groups, and committees.

Early research on social groups

Earlier research was concerned with whether small groups did better than individuals at various tasks (*e.g.*, factory work), while later research has been directed more toward the study of interaction patterns among members of such groups. In the method known as sociometry, members nominate others (*e.g.*, as best friends) to yield measures of preference and rejection in groups. Others have studied the effects of democratic and authoritarian leadership in groups and have greatly extended this work in industrial settings. In research on how people respond to group norms (*e.g.*, of morality or of behaviour), most conformity has been found to the norms of reference groups; *i.e.*, to such groups as families or close friends that are most important for people. The emergence and functioning of informal group hierarchies, the playing of social roles (*e.g.*, leader, follower, scapegoat), and cohesiveness (the level of attraction of members to the group) have all been extensively studied. Experiments have been done on processes of group problem solving and decision making, the social conditions that produce the best results, and the tendency for groups to make risky decisions. Statistical field studies of industrial work groups have sought the conditions for greatest production effectiveness and job satisfaction.

Social organizations. Such organizations as businesses and armies have been studied by social surveys, statistical field studies, field experiments, and laboratory experiments on replicas of their social hierarchies and communication networks. Although they yield the most direct evidence, field experiments present difficulties, since the leaders and members of such organizations may effectively resist the intervention of experimenters. Clearly, efforts to try out democratic methods in a dictatorship are likely to be severely punished. Investigators can study the effects of role conflict resulting from conflicting demands (*e.g.*, those from above and below) and topics such as communication patterns in social organizations. Researchers also have studied the sources of power and how it can be used and resisted. They consider the effectiveness of different organizational structures, studying variations in size, span of control, and the amount of power delegation and consultation. In factories, social psychologists study the effects of technology and the design of alternative work-flow systems. They investigate methods of bringing about organizational change; *e.g.*, in the direction of improving the social skills of people and introducing industrial democracy.

Changes in approach since 1900

Ways of looking at working organizations have changed considerably since 1900. Classical organization theory was criticized for its emphasis on social hierarchy, economic motivation, division of labour, and rigid and impersonal social relations. Later investigators emphasized the importance of flexibly organized groups, leadership skills, and job satisfaction based on less tangible rewards than salary alone. There has been a rather uneasy balance in the industrial social psychologist's concern with production and concern with people.

Personality. It is evident that there are individual differences in social behaviour; thus, people traditionally have been distinguished in terms of such personality traits as extroversion or dominance (see PERSONALITY). Some personality tests are used to predict how an individual is likely to behave in laboratory discussion groups, but usually the predictive efficiency is very small. Whether or not an individual becomes a leader of a group, for example, is found to depend very little on what such personality tests measure and more on that person's skills in handling the group task compared with the skills of others. Indeed, the same person may be a leader in some groups and a follower in others. Similar considerations apply to other aspects of social behaviour, such as conformity, persuasibility, and dependency. Although people usually perceive others as being consistent in exhibiting personality traits, the evidence indicates that each individual may behave very differently, depending on the social circumstances.

Socialization. The process by which personality is formed as the result of social influences is called socialization. Early research methods employed case studies of individuals and of individual societies (*e.g.*, primitive tribes). Later research has made statistical comparisons of numbers of persons or of different societies; differences in child-rearing methods from one society to another, for example, have been shown to be related to the subsequent behaviour of the infants when they become adults. Such statistical approaches are limited, since they fail to discern whether both the personality of the child and the child-rearing methods used by the parents are the result of inherited factors or whether the parents are affected by the behaviour of their children.

Problems in the process of socialization that have been studied by experimental methods include the analysis of mother–child interaction in infancy; the effects of parental patterns of behaviour on the development of intelligence, moral behaviour, mental health, delinquency, self-image, and other aspects of the personality of the child; the effects of birth order (*e.g.*, being the first-born or second-born child) on the individual; and changes of personality during adolescence. Investigators have also studied the origins and functioning of achievement motivation and other social drives (*e.g.*, as measured with personality tests).

Several theories have stimulated research into socialization; Freudian theory led to some of the earliest studies on such activities as oral and anal behaviour (*e.g.*, the effect of the toilet training of children on obsessional and other "anal" behaviour). Learning theory led to the study of the effects of rewards and punishments on simple social behaviour and was extended to more complex processes such as imitation and morality (*e.g.*, the analysis of conscience).

The self. Such concepts as self-esteem, self-image, and ego-involvement have been regarded by some social psychologists as useful, while others have regarded them as superfluous. There is a considerable amount of research on such topics as embarrassment and behaviour in front of audiences, in which self-image and self-esteem have been assessed by various self-rating methods. The origin of awareness of self has been studied in relation to the reactions of others and to the child's comparisons of himself with other children. Particular attention has been paid to the so-called identity crisis that is observed at various stages of life (*e.g.*, in adolescence) as the person struggles to discern the social role that best fits his self-concept.

Attitudes and beliefs. Research into the origins, dynamics, and changes of attitudes and beliefs has been carried out by laboratory experiments (studying relatively minor effects), by social surveys and other statistical field studies, by psychometric studies, and occasionally by field experiments. The origins of these socially important predispositions have been sought in the study of parental attitudes, group norms, social influence and propaganda, and in

Factors in the origin of predispositions of individuals

various aspects of personality. The influence of personality has been studied by correlating measured attitudes with individual personality traits and by clinical studies of cognitive and motivational processes; so-called authoritarian behaviour, for example, has been found to be deeply embedded in the personality of the individual. Early research based on statistical analyses of social attitudes revealed correlations with such factors as radicalism–conservatism. Later research on consistency provided extensive laboratory evidence of consistency but little evidence of it in actual political behaviour (*e.g.,* in attitudes on different political issues).

Research on attitude change has studied the effects of the mass media, the optimum design of persuasive messages, the effects of motivational arousal, and the role of opinion leaders (*e.g.,* teachers and ministers). Research has been carried out into the origins, functioning, and change of particular attitudes (*e.g.,* racial, international, political, and religious), each of which is affected by special factors. Attitudes toward racial minority groups, for example, are affected by social conditions, such as the local housing, employment, and the political situation; political attitudes are affected by social class and age; and religious attitudes and beliefs strongly reflect such factors as inner personality conflict.

Various specialties in social psychology. Many social psychologists are concerned with such aspects of public opinion (social survey) research as the design of standardized interviews and questionnaires. Forms of questions have been devised to compensate for errors that arise from the efforts to respond in a socially approved manner; some are designed to detect lying. Mass communications have been devised on the basis of research into persuasion. Use is also still made of Freudian symbolism and theory.

Research into the causes of mental disorders has shown the importance of social factors in the family and elsewhere. Mental patients often show deficiencies in social performance that may be the cause of other symptoms. Many social psychologists hold that social factors may also apply to such disorders as schizophrenia, which also seem to have hereditary and chemical bases. There has been a corresponding growth in the use of various kinds of social therapy in psychiatry (*e.g.,* group therapy, therapeutic communities, and social-skills training).

Considerable research has been devoted to industrial productivity, absenteeism, labour turnover, accidents, and job satisfaction. Factors that have been found to be important include the style of supervision and management, the size and composition of working groups, the technology and the work-flow systems, the span of control, and other features of the organizational structure. Research results point strongly toward the advantages of a less rigid hierarchical structure of authority, with more delegation of authority and consultation, training in supervisory skills, small and cooperative work teams, and interesting and varied work.

A major application of research in social interaction and group behaviour is in training in social skills, as in the T-groups, or sensitivity training, noted above. Role playing with video-tape playback and training in the imitation of other persons who serve as behavioral models are used in teaching people new skills. Actual training on the job has the advantage that there is no gap between the training and the work itself. All of these methods have been shown to be effective, depending on the job and the teacher. Social-skills training has been given successfully to industrial managers and supervisors, social workers and clergymen, interviewers, public speakers, mental patients, and juvenile delinquents.

A great deal of research has been done on factors underlying racial prejudice, but the understanding thus obtained has not had much effect upon the social problems involved. Similarly, the causes of delinquency and crime have been extensively studied, but it is not feasible to manipulate the factors influencing crime, such as genetic factors, methods of upbringing, and inequalities of opportunity. Social psychology has made some contribution to education; sociometry is quite widely practiced as a means of grouping children, and evidence is growing about the optimum styles of teacher behaviour. (M.Ar./Ed.)

Criminology. Criminology is the scientific study of the nonlegal aspects of crime (including juvenile delinquency). In its wider sense, embracing penology, it is thus the study of the causation, correction, and prevention of crime—seen from the viewpoints of such diverse disciplines as ethics, anthropology, biology, ethology (the study of character), psychology and psychiatry, sociology, and statistics.

Criminology and criminal law compared

Whereas the traditional legal approach to crime focusses on the action of crime and the protection of society, criminology focusses on the person of the criminal and the essential interests of the individuals of whom society consists. Whereas criminal law has been a relatively conservative force, often slow to change even where change has seemed imperative, criminology as a part of the developing social sciences of the past hundred years has been a revolutionary force—its object being not to replace the legal system in dealing with crime and punishment but to supplement it, making it less rigid and more sympathetic to approaches wider than strictly legal ones.

Without denying the value of "pure research," one must point to criminology and particularly penology as primarily practical subjects or "applied" disciplines. This practical value of criminological research can make itself felt in several ways. Its accumulated findings can give judges, prosecutors, lawyers, probation officers, and prison officials better understanding of crime and criminals, leading hopefully to more effective and humane sentencing and methods of treatment. Criminological research and knowledge can be equally at the disposal of legislators and administrators to assist in their task of reforming the law and improving penal and reformatory institutions. Essentially this purveyance of information represents a neutral role for criminologists; they garner the facts, and the various governmental officials decide for themselves what kind of practical conclusions to draw from the facts. Increasingly, however, some criminologists—like their counterparts in such fields as the atomic sciences—are demanding that scientists fully shoulder the moral and political responsibilities for their discoveries and for the use made of them instead of leaving vital decisions entirely to their governments. Thus some criminologists, for instance, insist upon actively campaigning against capital punishment, given the facts as they see them. Opponents of this activist role, on the other hand, contend that penological arguments are not sufficient but must be weighed along with political, social, religious, and moral arguments and that this all-round consideration should be left to responsible political bodies. The view does not deny the right of criminologists to express their opinions as ordinary citizens and voters; it does contend, nevertheless, that a government of officials responsive to the popular will, however fallible it may be, is less dangerous than a "government by experts."

The role of criminalistics and victimology

Another question involving the scope and functions of criminology is whether or not it should extend to the study of crime detection, involving such measures as photography, toxicology, fingerprint study, and the like. In several countries, notably Austria and Belgium, and at the school of criminology of the University of California at Berkeley, this so-called criminalistics has long been an important branch of criminological teaching and research, and the distinguished *Journal of Criminal Law, Criminology, and Police Science* (U.S.) devotes much of its space to criminal investigation. Actually, the only reason for excluding it from criminology is perhaps the expense of staff and equipment, which can be better borne by police colleges and similar specialized institutions. On the other hand, in recent decades criminology has undergone an important and perfectly legitimate extension of its territory by devoting much attention to so-called victimology—the study of the victim of crime, his relations to the criminal, and his role as a potential causal factor in crime.

Although the exclusion of criminalistics makes it easier to locate criminology on the map of scientific studies, its origin in, its close relations to, and its partial dependence on so many other disciplines result in considerable diversity and confusion regarding its proper place in the academic curriculum. Universities in continental Europe, when they do not ignore criminology altogether, tend to treat it as part of legal education; even where its principal

teachers are not lawyers. In Great Britain the only existing Institute of Criminology is part of the law faculty of Cambridge University; in other schools criminological research and teaching are usually divided between departments of sociology or social administration, law faculties, and institutes of psychiatry. In South America the anthropological and medical elements predominate, and in the United States, criminology, with a few notable exceptions, forms an established section of departments of sociology.

Given this situation in which criminology is submerged in other fields, it is not surprising that most teachers and researchers in criminology regard themselves first as sociologists, psychologists, lawyers, or whatever and only sec ondarily as criminologists. Their education contributes to this status; although a number may have pursued some criminological studies in their undergraduate years, criminology is largely a postgraduate discipline, at least in terms of major concentration for students.

This floating character of criminology weakens its position and tends to lend doubt to its claim to scientific status. Nevertheless, other disciplines—such as psychology, psychiatry, history, sociology, and social anthropology—have gone through similar birth pangs and, even after having achieved more or less assured positions, still face challenges to their claim to being scientific disciplines. The answer lies perhaps in historian H.R. Trevor-Roper's remark, "there are sciences and sciences." If the results of research can be viewed relatively, it is possible to perceive science in the criminologist's systematic application of sound research methods and his development of a body of facts from which he interprets general trends on a subject of real importance to mankind.

Historical development. The origins of criminology are generally dated from the late 18th century, when those imbued with a spirit of humanitarianism began questioning the cruelty, arbitrariness, and inefficiency of criminal justice and prison systems. From this period arose the so-called classical school of criminology, composed of such reformers as the Italian Marchese di Beccaria and the Englishmen Sir Samuel Romilly, John Howard, and Jeremy Bentham, all of whom may be said to have sought penological and legal reform rather than criminological knowledge per se—that is, knowledge about crime and criminals. Their principal aims were to mitigate legal penalties and subject judges to the principle of *nulla poena sine lege* or "due process of law" and also to reduce the application of capital punishment and humanize penal institutions. In all this they were moderately successful, but in their desire to make criminal justice "just," they tried to construct rather abstract and artificial equations between crimes and penalties, thereby forgetting the personal characteristics and needs of the individual criminal. Moreover, the object of punishment was seen as being primarily retribution, with deterrence occupying second place, and reformation lagging far behind.

Positivist school

By the second half of the 19th century these deficiencies, together with the influential teachings of the French sociologist Auguste Comte, had prepared the ground for the positivist school, which sought to bring a scientific neutrality into criminological studies. Instead of assuming a moral stance that focussed on measuring the criminal's "guilt" and "responsibility," the positivists attempted a morally neutral and social interpretation of crime and its treatment. Their leading figure, Cesare Lombroso (1836–1909), professor of psychiatry and anthropology at the University of Turin, sought through firsthand observation and measurement of prison inmates to determine the characteristics of criminal types. Some of his investigations led him into anthropometric interpretations—for example, his oft-criticized deduction of the "born criminal" with cranial, skeletal, and neurological malformations—but largely he and other positivists helped to introduce the ideas that crime has multiple causes and that most criminals are not born criminal but are shaped by their environmental upbringing and associations. With the positivists, therefore, the emphases in criminology had turned to experimental case studies and to preventive and rehabilitative measures. Without the upheaval caused by the positivists not only criminological research in the modern sense but also the present-day alternatives to capital punishment and old-fashioned imprisonment such as probation, suspended sentence, fines, and parole, inadequate as some of them are, would have been unthinkable.

Today, nevertheless, the feeling is widespread that the battle of ideas fought by the classical and positivist schools has not yet produced the secure foundations on which the criminology and the penal systems of the future can be built. Thus a third school, the postwar movement of "social defense," also originating in Italy, has tried to combine their best features and eliminate their excesses. This school disapproves of any rigid typology of criminals and stresses the uniqueness of human personality; it refuses the "scientism" of the positivists in favour of a strong belief in moral values—most importantly in balancing the rights of criminals and the rights of society. The school still speaks with too many voices, however, to be conclusively labelled.

Independent of these debates between schools, however, are the advances accomplished by such great figures as the statisticians Adolphe Quetelet (1796–1874) and André Michel Guerry (1802–66), the sociologists Gabriel Tarde (1843–1904) and Émile Durkheim (1858–1917), and of course Sigmund Freud, all of whom introduced a wealth of new ideas into the old problems of the social and individual characteristics of human, including criminal, behaviour.

Modern trends. The objectives of criminological research are sometimes said to be threefold: descriptive, causal, and normative. The descriptive aspect consists of the collection of relevant and reliable facts, together with their interpretation. The collection does not begin as a random and meaningless running after whatever phenomena happen to rouse the researcher's interest. Rather it is preceded by some hypothesis or "hunch," an assumption about what the researcher expects to find. The hypothesis "organizes" his inquiry. As the collection of facts proceeds, he may find either that his hypothesis is correct or that it requires revision or abandonment and thus the development of a new hypothesis to guide further research. The history of criminology reflects this perennial revision and renewal of inquiry, this continuous process of abandoning seemingly well-established theories in favour of new ones. Lombroso's theory of the "born criminal," the theory that all crime (or at least all economic crime) is due to poverty, and the theory that all juvenile delinquents come from broken homes had all to be drastically revised.

Forming and testing a hypothesis

The causal aspect involves relating the effects of one body of facts on another. Although regarded nowadays with some suspicion or indifference (even in the natural sciences), the search for causes is not being dispensed with altogether. So long as one does not jump to causal conclusions when arriving at statistical correlations or does not pressure "facts" into some proof of a popular theory, theories of causation can be useful in planning for the alleviation of crime and criminality.

The normative aspect, however, is more decidedly suspect. Research aimed at formulating so-called laws governing criminological phenomena has been thus far futile and does not look promising. What is sometimes regarded as a "law" has in reality been a mere trend. To diagnose statistical trends may be useful so long as their possible ephemeral nature is recognized, but trends are not laws, and there is thus but little scope for the normative approach in criminology.

Cross-disciplinary and cross-cultural approaches

Criminology, as suggested earlier, is cross-disciplinary and indeed draws methods or techniques from both the natural and the social sciences and must continually take heed of developments in other fields. It also depends increasingly on cross-cultural approaches; there have been recent statistical (though admittedly controversial) comparisons of "delinquent generations" in England, New Zealand, Denmark, and Poland; various studies of the sociological and statistical aspects of homicide; and studies of the ecological significance of "criminal areas" and of the possibility of predicting criminality.

In common with other disciplines, criminology must face such distinctions as between pure and applied research and between statistical and intuitive ways of thinking, but what

is almost unique to criminological research is its intense involvement with society and its difficulty in achieving "detachment." Not only do society's biases toward crime and punishment influence a criminologist's choice and execution of research but also he is dependent on the willing cooperation of governmental departments and other public authorities to secure essential raw material or data. There are only a few limited areas—such as adolescent delinquency and gang activities—in which research can be pursued privately without resort to official help.

Seen in the light of the previous remarks on the history of criminology, the development of criminological research can be divided into three stages: the prescientific, lacking any theoretical basis and largely identifiable with the work of the classical school; the semiscientific, possessing theories and hypotheses but without scientifically sound techniques and largely characteristic of the efforts of the positivist school; and the scientific, hopefully trying to repair earlier deficiencies and developing or improving the techniques described below.

Statistics. Often serving as the initial step in any research and regarded by some researchers, perhaps incorrectly, as the one and only reliable technique, the collection and interpretation of statistics for social and criminological purposes began in Europe early in the 19th century. The reputed "father" of this criminological method is the Belgian astronomer Adolphe Quetelet, who is perhaps best remembered for his famous "law," developed from the French and Belgian statistics, which showed that crime in any given country remains fairly constant over the long term (short-term fluctuations being insignificant). When he qualified himself, however, by saying that the volume and kind of crime were constant only so long as society's social, economic, and political conditions remained unchanged, he deprived the "law" of much of its significance for a rapidly changing modern world.

The manner and extent of data collection today differ considerably from country to country or, in federal unions like the United States, even from state to state or province to province. They differ in how often data are collected and published, in what items are given importance, in the choice between complete listings and sample surveys, in the ratio between governmental and private research, and so forth. Such far-reaching differences, together with the differences in law and its administration and in popular views and habits, have made it so far impossible to devise a meaningful system of international criminal statistics. Generally, however, there is increasingly less tendency to collect any and all data, regardless of reliability or practical value, and to concentrate on limited, reliable data involving matters of agreed upon importance.

Comparison of police and court statistics

A noteworthy distinction to be made is between police statistics and court statistics. Police data are nearer to the event but perhaps less reliable, and they usually describe the crimes only, not the criminals. The data from the courts, being based on convictions, do deal with the persons involved but include only the material brought forward by the prosecution and the defense. All criminal statistics indeed depend entirely on human factors such as the willingness of private individuals and officials to prosecute, on the popularity or unpopularity of the criminal laws at issue, and so forth. The figures also usually fail to rate very clearly the gravity of individual cases; except for such broad categories as "petty" and "grand," theft is theft regardless of the value of the objects stolen. Only recently have more detailed categories been attempted for criminological research.

Case studies. Also called the individual "case history," the case study concentrates on the career or life of one individual or group of individuals and is the method used primarily, though not exclusively, by psychologists, psychiatrists, and psychoanalysts. If well done, such histories can give deep insights into the personalities and motives of criminals, but the method does have shortcomings. Although the volume of case histories has grown large, their reliability is sometimes suspect—partly because of a criminal's natural reluctance to expose himself completely and partly because of the nature of the publication of case histories. Their publication is comparatively rare; professional ethics often forbid the exposure of details given confidentially, and those studies actually published may be too few to be typical and may even on occasion be designedly selective because of an investigator's wish to prove a theory.

Closely related to case studies are autobiographies and other books written by ex-prisoners, but in spite of their considerable human and scientific interest, they do suffer from even greater disadvantages, chiefly questionable objectivity. Sociologists have also contributed important studies of individuals in their social environments.

Typologies. The typological method involves classifying offenses, criminals, criminal associations, criminal areas, or whatever according to some criteria of relatedness or similarity. Thus there have been attempts to dichotomize criminals as either "normal" or "abnormal," "habitual" or "professional," or to form a continuum of criminals from the "insane" at one extreme through various career criminals, petty offenders, and white-collar criminals to "organized" or "professional" criminals at the other extreme. The typological method is less impersonal and heterogeneous than the statistical method and less individual or specific than the case study. Developed mainly in Germany and Austria and more recently in the United States, the method has been disputed; psychiatrists and psychoanalysts have especially questioned its value, primarily because it attempts to reduce complex phenomena to simple terms and tends to ignore important individual differences. Nevertheless, employed with restraint, the method is indispensable as a bridge between the two extremes, and it is in fact often used unknowingly by both statisticians and case students.

Experimental methods. A controlled experiment involves taking two closely related situations or groups, subjecting one of them to a specific change and comparing the subsequent characteristics of both. In the past, so-called experiments by judicial, penal, and reformatory institutions were not really controlled or even experimental in the scientific sense, for public agencies, at least in theory, are bound by the idea of justice to give equal treatment to equals, not one kind of treatment to one group and another kind to another group. Thus, generally speaking, most controlled experiments must be left to universities and other private bodies, and indeed the need for strict control and variable treatment has been recently accepted in such researches as Harvard University's Cambridge-Somerville Youth Study, which sought the effects of counselling on "pre-delinquent" boys.

Problems of crime prediction

Prediction studies. Criminological prediction—not unlike actuarial prediction used by insurance companies—is intended to forecast, usually in percentages, the future conduct of persons under certain conditions. Based on statistics or case histories or both, the predictions attempt to indicate probabilities—how any specific individuals or groups are likely to be affected by certain conditions or treatments. Thus, for example, various categories of criminals are listed as likely to be recidivistic.

The techniques involved in constructing prediction tables are too complicated to be discussed in brief; they have been developed and refined in the past 40 years mainly by Sheldon and Eleanor Glueck of Harvard University and also by several other authors in various countries. Statistical prediction by itself can never be conclusive; it must be subjected to rigorous validation for any individual or group, and even then it can merely show certain probabilities, which should be used in penal decisions only with the greatest caution and along with the lessons of experience derived from other sources. Nevertheless, the method can be valuable in supplementing the inevitably limited personal experience of judges and administrators, and indeed in recent decades prediction research has probably nowhere in the social sciences been more popular and urgent than in criminology.

Action research. Action research, which is often contrasted with experimental research, consists of drawing upon the observations of field workers and other persons directly involved with delinquents, potential delinquents, or prisoners. Thus, for example, have social workers attempted to help slum children and adolescents with their

problems and at the same time studied their delinquent behaviour, related it to their environment, and evaluated the results of the youth clubs or other services offered. The chief values of action research are that it aims at practical results through collaboration with fieldworkers, tries to build a bridge between theoretical and practical work, and may well dispense with formal hypotheses and simply aim at preventive tactics. Its best known and perhaps most successful example has so far been Clifford Shaw's Chicago Area Project, which, in close cooperation with the famous ecological studies of the University of Chicago, has tried to enlist suitable local people to deal with the social problems of their area.

Sociological research. Sociological research involves various methods—general surveys and personal interviews, as well as statistical, case-study, typological, experimental, and predictive techniques—and thus the purpose of classifying sociological research separately is chiefly to signal its focus or fields of interest. Mainly, criminology derives benefits from three fields of sociological study: (1) social institutions, involving such things as different conceptions of, or attitudes toward, property held by various societies or groups or the different effects of mass media on crime; (2) social groups, involving such things as the influence of juvenile gangs or criminal subcultures on individual criminal behaviour or the influence of prejudice on certain racial, national, or religious minorities; and (3) ecology, involving the study of different geographical areas and their rates and kinds of crime. Clifford Shaw's ecological studies of Chicago have been especially revealing in analyzing urban areas that either "breed" crime or "attract" crime. (H.M./Ed.)

Economics

No one has ever succeeded in neatly defining the scope of economics. Economists used to say, with Alfred Marshall, the great English economist, that economics is "a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing"—ignoring the fact that sociologists, psychologists, and anthropologists frequently study exactly the same phenomena. Another English economist, Lionel Robbins, has more recently defined economics as "the science which studies human behaviour as a relationship between (given) ends and scarce means which have alternative uses." This definition—that economics is the science of economizing—captures one of the striking characteristics of the economist's way of thinking but leaves out the macroeconomic approach to the subject, which is concerned with the economy as a whole.

What economists do

Difficult as it may be to define economics, it is not difficult to indicate the sort of questions that economists are concerned with. Among other things, they seek to analyze the forces determining prices—not only the prices of goods and services but the prices of the resources used to produce them. This means discovering what it is that governs the way in which men, machines, and land are combined in production and that determines how buyers and sellers are brought together in a functioning market. Prices of various things must be interrelated; how does such a "price system" or "market mechanism" hang together, and what are the conditions necessary for its survival?

These are questions in what is called "microeconomics," the part of economics that deals with the behaviour of such individuals as consumers, business firms, traders, and farmers. The other major branch of economics is "macroeconomics," in which the focus of attention is on aggregates: the level of income in the whole economy, the volume of total employment, the flow of total investment, and so forth. Here the economist is concerned with the forces determining the income of a nation or the level of total investment; he seeks to learn why full employment is so rarely attained and what public policies should be followed to achieve higher employment or more stability.

But these still do not exhaust the range of problems that economists consider. There is also the important field of "development economics," which examines the attitudes and institutions supporting economic activity as well as the process of development itself. The economist is concerned with the factors responsible for self-sustaining economic growth and with the extent to which these factors can be manipulated by public policy.

Cutting across these three major divisions in economics are the specialized fields of public finance, money and banking, international trade, labour economics, agricultural economics, industrial organization, and others. Economists may be asked to assess the effects of governmental measures such as taxes, minimum-wage laws, rent controls, tariffs, changes in interest rates, changes in the government budget, and so on.

The size of the profession

In the 19th century, economics was the hobby of gentlemen of leisure and the vocation of a few academicians; economists wrote about economic policy but were rarely consulted by legislatures before decisions were made. Today, there is hardly a government, international agency, or large corporation that does not have its resident economist. According to an estimate of the National Science Foundation (U.S.), for instance, there were 11,000 economists in the United States in 1966. Clearly, much depends on how one defines the job of an economist: the list of the National Science Foundation is confined to persons whose chief competence is in any one of the recognized economic specialities. Of the 11,000 professional economists, about 4,500 were employed as teachers of economics; the rest worked in various research or advisory capacities, either for themselves, for industry, or for government. This leaves out of account many others employed in accounting, commerce, marketing, and business administration; they may think of themselves as economists, but their professional expertise falls within other fields. There are perhaps another 10,000 economists in the rest of the world—their numbers have never been counted. It would be reasonable to estimate the total number of professional economists in the world in 1970 at 20,000, a number that was apparently growing at about 5 percent per year. There were about 75 English-language journals in economics and another 25 in various foreign languages, with new ones appearing every year. This implies the publication of about 1,500 scientific papers per year, not to mention the 700 new books on economics published every year. This is indeed "the age of economists," and the demand for their services seems insatiable.

HISTORICAL DEVELOPMENT OF ECONOMICS

The work of Adam Smith

The effective birth of economics as a separate discipline may be traced to the year 1776, when the Scottish philosopher Adam Smith published *An Inquiry into the Nature and Causes of the Wealth of Nations.* There was, of course, economics before Adam Smith: the Greeks made significant contributions and so did the medieval scholastics; from the 15th to the 18th century, an enormous pamphlet literature appeared that developed the implications of economic nationalism, a body of thought now known as "mercantilism"; for a brief period in the 18th century the French "physiocrats" developed a fairly sophisticated economic model; and several other 18th-century figures can compete with Smith for the title of "first economist." Nevertheless, Adam Smith wrote the first full-scale treatise on economics and, by his magisterial influence, founded what later generations were to call the "English School of Classical Political Economy."

Analysis of the market. The *Wealth of Nations,* as its title suggests, is essentially a book about economic development and about policies that promote or hinder development. In its practical aspects it is an attack on the protectionist doctrines of the mercantilists and a brief for free trade. But in the course of attacking "false doctrines of political economy," Adam Smith was led to analyze the workings of a free-enterprise system as a governor of human activity. In a competitive market each individual, being one among many, can exert only a negligible influence on prices; each must take prices as they come and is free only to vary the quantities bought and sold at given prices; yet the sum of all individuals' separate actions determines prices. The "invisible hand" of the market, as Adam Smith was fond of saying, assures a

social result that is independent of individual intentions and thus creates the possibility of an objective science of economic behaviour. Adam Smith believed that he had found, in the competitive market, an instrument capable of converting "private vices" (like selfishness) into "public virtues" (like maximum production). But this is only true if the competitive system is embedded in an appropriate legal and institutional framework, an insight that Adam Smith developed at length but that was largely forgotten by later generations. Within this great tome on the theme of rich and poor nations was contained a simple theory of value (or prices), a crude theory of distribution, an even cruder theory of international trade, and a primitive theory of money; but with all their imperfections, these were the building blocks of classical and modern economics. The book's very fecundity gave it strength because it left so much for disciples to tidy up.

The later classical economists

Construction of a system. David Ricardo's *Principles of Political Economy and Taxation* (1817) was, in one sense, simply a critical commentary on the *Wealth of Nations;* in another sense, it gave an entirely new twist to the developing science of political economy. Ricardo invented the concept of the "economic model," a tightly knit logical apparatus consisting of a few strategic variables, an apparatus that was capable of yielding, after a bit of manipulation, results of enormous practical import. At the heart of the Ricardian system is the notion that economic growth must sooner or later be arrested, owing to the rising cost of growing food on a limited land area. An essential ingredient of this argument is the Malthusian principle—enunciated in Thomas Malthus' *Essay on Population* (1798)—that population tends to increase up to the limits set by the existing supply of food, thus holding down wages. As the labour force increases, extra food to feed the extra mouths can be produced only by extending cultivation to less fertile soil or by applying capital and labour to land already under cultivation (with diminishing results because of the so-called law of diminishing returns). Although wages are held down, profits do not rise proportionately because tenant farmers outbid each other for superior land. The chief beneficiaries of economic progress, therefore, are the landowners.

Since the root of the trouble, according to Ricardo, is the declining yield of wheat per unit of land, one obvious solution is to import cheap wheat from other countries. Eager to show that Britain would benefit from specializing in manufactured goods and exporting them in return for food, Ricardo hit upon the "law of comparative costs" as proof. He assumed that, within countries, labour and capital are free to move in search of the highest returns; between countries, however, they are not. In these circumstances, Ricardo showed, the benefits of trade are determined by a comparison of costs *within* each country, rather than by a comparison of costs *between* countries. It pays a country to specialize in the production of those goods that it can produce *relatively* more efficiently and to import everything else; although India may be able to produce everything more efficiently than England, India is nevertheless well advised to concentrate its resources on textiles, in which its efficiency is relatively greater, and to import British capital goods. The beauty of the argument is that if all countries take full advantage of the territorial division of labour, total world output is certain to be larger than it will be if some or all countries try to become self-sufficient. Ricardo's law became the fountainhead of 19th-century free-trade doctrine, which would have been enough, if he had said nothing else, to give him a place in the economists' pantheon.

Ricardo's influence

The influence of Ricardo's treatise was felt almost as soon as it was published, and for over half a century the Ricardian system dominated economic thinking in Britain. In 1848 John Stuart Mill's restatement of Ricardo's thought in his *Principles of Political Economy* brought it new authority. After 1870, however, most economists turned their backs on the range of problems that had concerned Ricardo and began to re-examine the foundations of the theory of value; that is, they became interested in the theory of why goods exchange at particular prices, so that for a while they devoted almost all of their efforts to the problem of resource allocation under conditions of perfect competition.

Marxism. A few words must first be said, however, about the last of the classical economists, Karl Marx. The first volume of *Das Kapital* appeared in 1867; the second and third after his death, in 1883 and 1894. For a generation, therefore, the competitive market theorists jostled with the followers of Marx. By 1900 the intellectual battle was over, and thereafter professional economists largely lost interest in Marx. Despite the Russian Revolution, despite what amounts to official endorsement of Marxism in one-third of the world, and despite the lingering influence of Marx's ideas, Marxian economics has been moribund ever since Marx's death in 1881. If Marx may be called "the last of the classical economists," it is because to a large extent he found his economics not in the real world but in the teachings of Smith and Ricardo. They had espoused a "labour theory of value," which holds that products exchange roughly in proportion to the labour costs incurred in producing them. Marx worked out all the logical implications of this theory and added to it "the theory of surplus value," which rests on the axiom that human labour alone creates all value and hence constitutes the sole source of profits. It is an axiom in the sense that it cannot be established in terms of the theory itself: it must be imported from without. To say that an economist is a Marxian economist is in effect to say that he shares the value judgment that it is socially undesirable for some people in the community to derive their income merely from the ownership of property. Since few professional economists in the 19th century accepted this ethical postulate and most were indeed inclined to find some social justification for the existence of private property and the income derived from it, Marxian economics fell on deaf ears. The Marxian system, moreover, culminated in three great generalizations: the tendency of the rate of profit to fall, the growing impoverishment of the working class, and the increasing severity of business cycles, of which the first is the linchpin of all the others. Marx's exposition of the "law of the declining rate of profit" is invalid; with it all of Marx's other predictions fall to the ground. In addition, Marxian economics had little to say on some of the practical problems that are the bread and butter of economists in any society. This is enough to suggest why Marxian economics failed to make many converts among academic economists. Marxists will reply that the reason is simply that academic economists have always been "lackeys of the capitalist class." Perhaps so, but the fact remains that Marx has had virtually no effect on modern economic thought.

The three generalizations of Marxism

The marginalists. The marginal revolution was essentially the work of three men: Stanley Jevons, an Englishman; Carl Menger, an Austrian; and Léon Walras, a Frenchman. Their contribution was the replacement of the labour theory of value by the marginal utility theory of value; their explanation of prices began with the behaviour of consumers in choosing among increments of goods and services (see ECONOMIC THEORY). The idea of emphasizing the marginal or last unit proved in the long run to be more significant than the introduction of utility. It was the consistent application of marginalism that marks the true dividing line between classical theory and modern economics. The classical political economists saw the economic problem as that of predicting the effects of changes in the quantity of capital and labour on the rate of growth of national output. The marginal approach, however, focussed on the conditions under which these factors tend to be allocated with optimal results among competing uses—optimal in the sense of maximizing consumers' satisfactions.

Throughout the last three decades of the 19th century, the English, Austrian, and French contributors to the marginal revolution largely went their own way. The Austrian school dwelt on the importance of utility as the determinant of value and vehemently attacked the classical economists as completely outmoded. A brilliant second-generation Austrian economist, Eugen von Böhm-Bawerk, applied the new ideas to the determination of the rate of interest, putting his stamp for all time on capital theory

(see ECONOMIC THEORY). The English school, led by Alfred Marshall, sought a reconciliation with the doctrines of the classical writers. The classical authors, Marshall argued, concentrated their efforts on the supply side in the market; marginal utility theory was concerned with the demand side, but prices are determined by both supply and demand, just as a pair of scissors cuts with both blades. Marshall, seeking to be practical, applied his "partial equilibrium analysis" to particular markets and industries.

The leading French marginalist was Léon Walras, who carried the approach furthest by describing the economic system in general mathematical terms. For each product there is a "demand function" that expresses the quantities of the product that consumers demand as depending on its price, the prices of other related goods, the consumers' incomes, and their tastes. For each product there is also a "supply function" that expresses the quantities producers will supply as depending on their costs of production, the prices of productive services, and the level of technical knowledge. In the market, for each product there is a point of "equilibrium"—analogous to the equilibrium of forces in classical mechanics—at which a single price will satisfy both consumers and producers. It is not difficult to analyze the conditions under which equilibrium is possible for a single product. But equilibrium in one market depends on what happens in other markets (a "market" in this sense being not a place or location but a complex of transactions involving a single good), and this is true of every market. There are literally millions of markets in a modern economy, and therefore "general equilibrium" involves the simultaneous determination of partial equilibria in all markets. Walras' efforts to describe the economy in this way led the historian of economic thought Joseph Schumpeter to call his work "the Magna Carta of economics." Walrasian economics is undeniably abstract, but it provides an analytical framework for incorporating all of the elements of a complete theory of the economic system. It is not too much to say that nearly the whole of modern economics is Walrasian economics. Certainly, modern theories of money, of employment, of international trade, and of economic growth are all Walrasian general equilibrium theories in a simplified form.

The neoclassical era

The years between the publication of Marshall's *Principles of Economics* (1890) and the Great Crash in 1929 may be described as years of reconciliation, consolidation, and refinement. The three national schools gradually coalesced into a single mainstream. The theory of utility was reduced to an axiomatic system that could be applied to the analysis of consumer behaviour under various circumstances, such as a change in income or price. The concept of marginalism in consumption led eventually to the idea of marginal productivity in production, and with it came a new theory of distribution in which wages, profits, interest, and rent were all shown to depend on the "marginal value product" of a factor. Marshall's concept of "external economies and diseconomies" was developed by his leading pupil, Arthur Pigou, into a far-reaching distinction between private costs and social costs, thus laying the basis of welfare theory as a separate branch of economic inquiry. There was a gradual development of monetary theory, which explains how the level of all prices is determined as distinct from the determination of individual prices, notably by the Swedish economist Knut Wicksell. In the 1930s the growing harmony and unity of economics was rudely shattered, first by the simultaneous publication of Edward Chamberlin's *Theory of Monopolistic Competition* and Joan Robinson's *Economics of Imperfect Competition* in 1933 and then by the appearance of John Maynard Keynes's *General Theory of Employment, Interest and Money* in 1936.

The critics. Before going on, it is necessary to take note of the rise and fall of the German Historical school and the American Institutionalist school, which levelled a steady barrage of critical attacks on the orthodox mainstream. The German historical economists, who had many different views, basically rejected the idea of an abstract economics with its supposedly universal laws; they urged the necessity of studying concrete facts in national contexts. While they gave impetus to the study of economic history, they failed to persuade their colleagues that their method was invariably superior. The institutionalists are more difficult to categorize. "Institutional economics," as the term is narrowly understood, refers to a movement in American economic thought associated with such names as Thorstein Veblen, Wesley Clair Mitchell, and John R. Commons. These writers had little in common aside from their dissatisfaction with the abstract theorizing of orthodox economics, its tendency to cut itself off from the other social sciences, and its preoccupation with the automatic market mechanism. They failed to develop a theoretical apparatus that would replace or supplement the orthodox theory. This may explain why the phrase "institutional economics" has become little more than a synonym for "descriptive economics." The hope that institutional economics would furnish a new interdisciplinary social science proved stillborn. (This is perhaps not surprising, because it was by abstracting purely economic forces from the totality of social interactions that economics got so far ahead of the other social sciences in theoretical rigour.) Although there is no longer an institutionalist movement in economics, the spirit of institutionalism is alive in such works as the Harvard economist John Kenneth Galbraith's *The Affluent Society* (2nd ed., 1969) and *The New Industrial State* (1967).

Returning to the innovations of the 1930s, the theory of monopolistic or imperfect competition remains somewhat controversial to this day. The older economists had devoted all their attention to two extreme types of market structure, that of "pure monopoly," in which a single seller controlled the entire market for one product, and that of "pure competition," characterized by many sellers, highly informed buyers, and a single, standard product. The theory of monopolistic competition gave recognition to the range of market structures that lie between these extremes, including (1) markets having many sellers with "differentiated products," employing brand names, guarantees, and special packaging that cause consumers to regard the product of each seller as unique; (2) "oligopoly," markets dominated by a few large firms; and (3) "monopsony," markets with a single monopolistic buyer and many sellers (see ECONOMIC THEORY). The theory produced the powerful conclusion that competitive industries in which each seller has a partial monopoly because of product differentiation will tend to have an excessive number of firms, all charging a higher price than they would if the industry were perfectly competitive. Since product differentiation—and the associated phenomenon of advertising—seems to be characteristic of most industries in developed capitalist economies, the new theory was immediately hailed as injecting a healthy dose of realism into orthodox price theory. Unfortunately, its scope was not great enough. It failed to provide a satisfactory theory of price determination under conditions of oligopoly. In advanced economies many of the manufacturing industries are oligopolistic. The result has been to leave a somewhat undigested lump at the centre of modern price theory, a constant reminder of the fact that economists still lack an adequate explanation of the conditions under which the giant firms of rich countries conduct their affairs.

Forms of imperfect competition

Keynesian economics. The second major breakthrough of the 1930s, the theory of income determination, was primarily the work of one man—John Maynard Keynes. Keynes asked questions that in some sense had never been asked before; he was interested in the level of national income and the volume of employment rather than in the equilibrium of the firm or the allocation of resources. It was still a problem of demand and supply, but "demand" here means the total level of effective demand in the economy, and "supply" means the nation's capacity to produce. When effective demand falls short of productive capacity, the result is unemployment and depression; when it exceeds the capacity to produce, the result is inflation. The heart of Keynesian economics consists of an analysis of the determinants of effective demand. If one ignores foreign trade, effective demand consists essentially of three spending streams: consumption expenditures, investment expenditures, and government expenditures, each of which is independently determined. Keynes attempted to show

that the level of effective demand so determined may well exceed or fall short of the physical capacity to produce goods and services: that there is no automatic tendency to produce at a level that results in the full employment of all available men and machines. This fundamental implication of the theory came as something of a shock to exponents of the traditional economics who had been inclined to take refuge in the assumption that economic systems tend automatically to full employment. By keeping his attention focussed on macroeconomic aggregates, like total consumption and total investment, and by a deliberate simplification of the relations between these economic variables, Keynes achieved a powerful model that could be applied to a wide range of practical problems. His system subsequently underwent considerable refinement—some have said that Keynes himself would hardly have recognized it—and became thoroughly assimilated into the body of received doctrine (see ECONOMIC THEORY). Still, it is not too much to say that Keynes is perhaps the only economist to have added something really new to economics since Walras and perhaps since Ricardo.

Keynesian economics as conceived by Keynes was entirely "static"; that is, it did not involve time as an important variable. But a disciple of Keynes, Roy Harrod, soon developed a simple macroeconomic model of a growing economy; in 1948 he published *Towards a Dynamic Economics*, launching an entirely new speciality, "growth theory," which absorbed the attention of an increasing number of economists.

Postwar developments. In the 25-year period following World War II, economics was so totally transformed that those who studied it before the war might as well have lived in another world. First of all, there was an enormous increase in the use of mathematics, which came to permeate virtually every branch of economics. Previously, few economists had made use of mathematics other than differential and integral calculus. Matrix algebra became important with the advent of "input–output analysis," an empirical method of reducing the technical relations between industries to a manageable system of simultaneous equations; it was an attempt to put quantitative flesh on the bones of a general equilibrium model of the economy. A closely related phenomenon was the development of "linear programming" and "activity analysis," which opened up a whole host of industrial problems to numerical solution and introduced economists for the first time to the mathematics of "inequalities" rather than exact equations. Likewise, the emergence of "growth economics" promoted the use of difference and differential equations.

New mathematical tools

Hand in hand with the spread of mathematical economics went an increasing sophistication of empirical work under the rubric of "econometrics," comprising a combination of economic theory, mathematical model building, and statistical testing of economic predictions. The development of econometrics had an impact on economics in general, since those who formulated new theories began to cast them in terms that allowed them to be empirically tested.

The postwar years also saw a renewal of interest in the underdeveloped countries. Economists became aware that they had too long neglected "an inquiry into the causes of the wealth of nations." There was also a conviction that economic planning of one variety or another was needed to close the gap between the rich and poor countries. Out of these concerns came the field of development economics. Regional economics, urban economics, health economics, and the economics of education are other offshoots of the mainstream since 1945.

The postwar tendencies in economic thought were best exemplified, not by the emergence of new techniques or by the addition of new parts to the economics curriculum but by the disappearance of divisive "schools," by the increasingly standardized professional training of economists all over the world, and by the transformation of the science from a rarefied academic exercise to an operational discipline geared to practical advice. This transformation brought prestige to the profession but also a new responsibility: now that economics really mattered, economics had to reckon with the conflict that so often exists between analytical precision and economic relevance.

The question of relevance was at the centre of a "radical critique" of economics that developed along with the campus revolts of the late 1960s. The radical critics declared that economics had become a defense of the status quo and that its practitioners had joined the power elite. The marginal techniques of the economists, ran the argument, were profoundly conservative in their bias, and encouraged a piecemeal rather than a revolutionary approach to social problems; likewise, the tendency in theoretical work to ignore the everyday context of economic activity amounted in practice to the tacit acceptance of prevailing institutions. The critics said that economics should abandon its claim of being a value-free social science and address itself to the great questions of the day—those of civil rights, poverty, imperialism, and nuclear war—even at the cost of analytical rigour and theoretical elegance.

It is true that the study of economics encourages a belief in reform rather than revolution; economics as a science does not provide enough certitude for any thoroughgoing reconstruction of the social order. It is also true that most economists tend to be deeply suspicious of monopoly in all forms, including state monopolies, and to favour competition between independent producers as a way of diffusing economic power. Finally, most economists prefer to be silent on large questions if they have nothing to offer beyond the expression of personal preferences: most economists as economists are fundamentally concerned with the professional standards of their discipline, and this may mean in some cases frankly conceding that economics has as yet nothing very interesting to say about these questions. (It does not mean, however, that they desire to justify the status quo.) Yet the radical critique of modern economics was not to be lightly dismissed. The radical economists were posing issues that were important. At the very least it could do the economic researcher no harm to think about the social and political relevance of his work.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY ECONOMICS

The role of value judgments

Economists are sometimes confronted with the charge that their discipline is not a science. Human behaviour, it is said, cannot be analyzed with the same objectivity as the behaviour of atoms and molecules. Value judgments, philosophical preconceptions, and ideological biases must interfere with the attempt to derive conclusions that are independent of the particular economist espousing them. Moreover, there is no laboratory in which economists can test their hypotheses.

This argument raises issues for all of the social sciences. Only a very general reply can be given here. Economists are wont to distinguish between "positive economics" and "normative economics." Positive economics seeks to establish facts: Will a subsidy to butter producers lower the price of butter? Will a rise in wages in the automobile industry reduce the employment of automobile workers? Will devaluation improve the balance of payments? Does monopoly foster technical progress? Normative economics, on the other hand, is not concerned with matters of fact but with questions of policy, of "good" or "bad": Should the goal of price stability be sacrificed to that of full employment? Should income be taxed at a progressive rate? Should there be legislation in favour of competition?

Positive economics in principle involves no judgments of value; its findings may be as impersonal as those of astronomy and meteorology, two natural sciences that are also denied the advantage of conducting laboratory experiments. As the British philosopher David Hume argued 200 years ago, there is no logical way to deduce "ought" from "is" or prescriptions from descriptions; all statements of fact are ethically neutral. In that sense a value-free economics is possible (at least in principle): if economics is about the application of means to achieve given ends, there would seem to be no reason why one cannot analyze the allocation of means to achieve *any* end. This is not to deny that most of the interesting economic propositions involve the addition of definite value judgments to a body of established facts, that ideological bias creeps into the very selection of the questions that economists investigate, that what is a means from one point of view may be an

end from another, nor even that much practical economic advice is loaded with concealed value judgments, the better to persuade rather than merely to advise. This is only to say that economists are human. The commitment of economists the world over to the ideal of value-free positive economics (or to the candid declaration of personal values in normative economics) serves as a defense against the attempts of special interests to bend the science to their own purposes. The best assurance against bias on the part of any particular economist is the criticism of other economists. The best protection against special pleading in the name of science is the professional standards of scientists.

Methods of inference. But how, one may ask, are facts established in a science that cannot conduct experiments?

Importance of statistics

In essence, the answer is: by means of statistical inference. Economists typically begin by describing the area under study according to what they feel to be important. Then they construct a "model" of the real world, deliberately repressing some of its features and emphasizing others; they abstract, isolate, and simplify, thus imposing order on a world that at first glance appeared disorderly. Having evolved an admittedly unrealistic representation of the real world, they then manipulate the model by a process of logical deduction, arriving eventually at some prediction or implication that is of general significance. At this point, they return to the real world to see whether or not the prediction is borne out by observed events.

But the observable events that are available to test a theory never exhaust the population of all such events: they are merely a sample of it. This raises the problem of statistical inference; namely, what can be inferred about a population from a sample of the population? The theory of statistical inference is simply an agreed-upon procedure for making such inferences, but in the nature of the case it never succeeds in removing all elements of judgment from an inference. Thus the empirical truths of economics are invariably surrounded by a band of doubt, and economists speak of them as "probable" or "likely"; they are propositions in which economists have "a certain degree of confidence" because it is unlikely that they could have come about by chance.

It follows that judgments are at the heart of both positive and normative economics. It is easy to see, however, that judgments about "degrees of confidence" and "statistical levels of significance" are of a totally different order from those that crop up in normative economics. When men say that every individual should be allowed to spend his income as he likes, that people should not be free to control material resources and to employ others, or that governments must offer relief for the victims of inexorable economic forces, they are making the kind of value judgments that laymen have in mind when they accuse economists of producing personal preferences in the guise of scientific conclusions. There is no room for such value judgments in positive economics.

Testing theories. In the past some economists tended to claim too much for their propositions. Economic models were said to be based on fundamental axioms and premises about economic behaviour that were absolutely true a priori because they were derived from an examination of one's own economic behaviour. Since the theorems of the model were deduced from these axioms by the laws of logic, the theorems also were held to be true a priori. Economic models did not need to be confronted with empirical evidence.

This extreme apriorist position may be contrasted with the ultraempiricist view, which holds that one must begin and end with observable facts; the latter approach, however, has never appealed to more than a small minority of economists. In the middle ground between these two sharply opposing views is the methodological position that has found increasing acceptance among modern economists. It argues that one must test the predictions or conclusions of a model but without worrying too much about the realism of its premises, axioms, or assumptions. Most assumptions in economic theory cannot be tested directly. For example, there is the famous assumption of price theory that businessmen strive to maximize profits.

The problem of validity

Attempts to find out whether they do, by asking them, usually fail; after all, businessmen are no more fully conscious of their own motives than other people are. A logical approach would be to observe businessmen in action. But that would require knowing what sort of action is associated with profit maximizing, which is to say that one would have drawn out all the implications of a profit maximizing model. Thus one would be testing an assumption about business behaviour by comparing the predictions of a theory of the firm with observations from the real world.

This is not as easy as it sounds. Since the predictions of economics partake of the nature of probability statements, there can be no such thing as a conclusive, once-and-for-all test of an economic hypothesis. The science of statistics cannot prove any hypothesis, it can only fail to disprove it. A theory that survives a statistical test is not true as such; it is only provisionally true because it has so far resisted all attempts to falsify it. Attempts to falsify economic hypotheses never yield unambiguous results, and hence economic theories tend to survive until they are falsified repeatedly with new or better data. This is not because they are *economic* theories but because the attempt to compare predictions with outcomes in the social sciences is always limited by the rules of statistical inference.

It is not remarkable that competing theories exist to explain the same phenomena, with economists disagreeing as to which theory is to be preferred. While virtually all economists today agree that theories should be submitted to empirical testing and that the theory is to be preferred that allows predictions that conform, in a probabilistic sense, most closely to observable events, this precept can be very difficult to apply in practice. There have been periods in the history of economics when there was overwhelming agreement in the profession as to which models or theories were "true." But a period of consensus may be followed by a generation of doubt until a new departure is made that succeeds in producing a new consensus. In this, economics is not very different from physics.

Much has been written about the doubtful accuracy of economists' predictions. Of course, economists cannot foretell the future as such; only soothsayers do that. Economists can foretell the effects of specific changes in the economy, but they are better at predicting the direction than the actual magnitude of events. When economists predict that a tax cut will raise national income, one may be confident that the prediction is accurate; when they predict that it will raise national income by a certain amount in three years, however, the forecast is likely to miss the mark. The reason is that most economic models do not contain any explicit reference to the passage of time and hence have little to say about how long it takes for a certain effect to make itself felt. Short-period predictions generally fare better than long-period ones, in part because most economic models rely on propositions about the plans and intentions of economic agents, whereas the data on which the theories are tested are derived from past events. This is disappointing, but it does not mean that economics is not a science.

Microeconomics. Since Keynes, economic theory has been of two kinds: macroeconomics—or the study of the determinants of national income—and the traditional microeconomics. The latter approaches the economy as if it were made up only of business firms and households (ignoring governments, banks, charities, trade unions, and all other economic institutions) interacting in two kinds of markets—product markets and markets for productive services, or factor markets. Households appear as buyers in product markets and as sellers in factor markets, where they offer men, machines, and land for sale or hire. Firms appear as sellers in product markets and as buyers in factor markets. In each type of market, price is determined by the interaction of demand and supply, and the problem of microeconomic theory is to say something meaningful about the forces that make up demand and supply.

The firm and the household

Theory of choice. At first it appears that all one can say is that everything depends on everything else. But firms and households do not behave in a vacuum. Firms face certain technical constraints in producing goods and ser-

vices, and households have definite preferences for some products over others. It is possible to express the technical constraints facing business firms by writing down a series of "production functions," one for each firm. A production function is simply a kind of equation that expresses the fact that the output of a firm depends on the quantity of inputs it employs and, in particular, that inputs can be technically combined in different proportions to produce a given level of output. A production engineer could calculate, on the basis of existing technical knowledge, the largest possible output that could be produced with every possible combination of inputs and in this way could define a boundary to the range of production possibilities open to a firm. By itself this does not tell how much the firm will produce or what mixture of products it will make or what combination of inputs it will adopt: these depend on the prices of products and the prices of inputs (or "factors of production"), which have yet to be determined. One may assume that the firm is motivated in a particular way: it wants to maximize profits, which are defined as the difference between the sales value of its output and the money outlays required to obtain its inputs. It will, therefore, select that combination of inputs that minimizes the costs of producing any given quantity of output and will select from the range of possible combinations of products that combination that maximizes its revenues. This is to say that it always tries to move along its production function, along the edge of the boundary of technical possibilities. But where it ends depends, in part, on the demand for its products. This leads to the part played by households in the system.

Households are endowed with definite "tastes" that can be expressed in a series of "utility functions," one for each household. A utility function is an equation like a production function, expressing the fact that the pleasure or satisfaction that households derive from consumption depends on the products that they purchase and on the various ways in which they combine these products in consumption to yield a given level of satisfaction. The utility function need not be specified in the same detail as a production function. One may think of it as a general description of the household's preferences between all the paired alternatives with which it will be confronted. Here, too, it is necessary to assume something about motivation to make any progress: the assumption is that households seek to maximize satisfaction, distributing their given incomes among available consumer goods in such a way as to derive the largest possible "utility" from consumption. Their incomes, however, remain to be determined.

Supply and demand curves

The purpose of production functions in economic theory is to provide an anchor in the bedrock of technology from which to derive the "supply curves" of firms in product markets and the "demand curves" of firms in factor markets. Similarly, the purpose of utility functions is to provide an anchor in subjective "tastes" from which to derive the "demand curves" of households in product markets and the "supply curves" of households in factor markets. All of these demand and supply curves express the quantities demanded and supplied as a function of prices, not because price is the only determinant of economic behaviour but because the purpose is to have a theory of price determination. Much of economic theory is devoted to showing how various production and utility functions, coupled with certain assumptions about behaviour, lead to demand and supply curves in which the quantity demanded is inversely related and the quantity supplied positively related to price. The figure depicts these relationships (curves would be just as suitable as straight lines).

Not all demand and supply curves look alike. The essential point is that most demand curves are negatively inclined, while most supply curves are positively inclined. This may seem a modest result for a great deal of effort, but the argument has powerful implications. The participants in a market will be driven automatically to the price at which the two curves intersect; this price p is called the "equilibrium" price or "market-clearing" price because it is the only price at which supply and demand are equal. If it is a market for butter, any change in the production

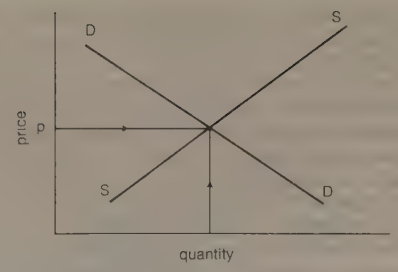


Illustration of the relationship of price to supply (S) and demand (D).

function of dairy farmers or in the utility function of butter consumers or in the prices of cows, grassland, and milking equipment or in the incomes of butter consumers or in the prices of nondairy products that consumers buy can be shown to lead to definite changes in the equilibrium prices of butter and in the equilibrium quantity of butter produced. Better still, the effects of a government ceiling on the price of butter or of a tax on butter producers or of a price-support program for dairy farmers can be predicted with almost perfect certainty. As a rule, the prediction will refer only to the direction of change (the price will go up or down); but if the demand and supply curves of butter can be defined in quantitative terms on the basis of past data, one may be able to predict the actual magnitude of the change.

Theory of allocation. This analysis of the behaviour of firms and households is to some extent symmetrical: all economic agents are conceived of as ordering a series of attainable positions in terms of an entity they are trying to maximize. For a firm these attainable positions are essentially input combinations; for a household they are product combinations. From the maximizing point of view, some combinations are better than others; the best combination is called the "optimal" or "efficient" combination. The rule for efficient, optimum allocation may now be stated baldly: an optimum allocation is one that equalizes the returns of the marginal or last unit to be transferred between all the possible uses. In the theory of the firm, an optimum allocation of outlays among the factors of production implies that the "marginal physical product" of an additional dollar devoted to hiring the services of any one of the factors is the same for all factors; the so-called law of eventually diminishing marginal productivity, a property of a wide range of production functions, ensures that such an optimum exists. In the theory of consumer behaviour an optimum situation obtains when the consumer has distributed his given income in such a way that the "marginal utility" of each additional dollar spent on any of the products purchased is equal for all products; the "law of eventually diminishing marginal utility," a property of a wide range of utility functions, ensures that such an optimum exists. These are merely particular examples of the "equimarginal principle," which is not only at the core of the theory of the firm and the theory of consumer behaviour but also underlies the theory of money, of capital, and of international trade. In fact, the whole of microeconomics is nothing more than the spelling out of this principle in ever wider contexts.

Law of diminishing marginal productivity

The equimarginal principle is, of course, applicable to any decision that involves alternative courses of action. Economics furnishes a technique for thinking about decisions, whatever their character and whosoever makes them. Military planners may, for example, consider a variety of weapons in the light of a single objective, that of damaging an enemy; some of the weapons are effective against the enemy's army, some against the enemy's navy, and some against his air force; the problem is to find an optimal allocation of the defense budget, one that equalizes the marginal contribution of each type of weapon. But defense departments rarely have single objectives; along with maximizing damage to an enemy there may be another objective, such as minimizing losses from attacks. In that case, more than the equimarginal principle is needed for a decision; it is necessary to know how the department ranks the two objectives in order of importance,

since different rankings will imply different optima. But a ranking of objectives is simply a utility function or a preference function.

In other words, when an institution pursues multiple ends, decisions about how to achieve them require a weighting of the ends. Every decision involves a "production function"—a statement of what is technically feasible—and a "utility function"; the equimarginal principle is then invoked to provide an efficient, optimal strategy. This applies just as well to the running of hospitals, churches, and schools as to the conduct of a business enterprise, to the location of an international airport as well as to the design of a development plan for an underdeveloped country. This is why economists crop up in what seem to be the most unlikely places, advising on activities that are obviously not being conducted for economic reasons.

The different approach of macroeconomics

Macroeconomics. There is, however, an approach to economics in which the foregoing considerations do not apply. That is the field known as macroeconomics. In macroeconomics one is concerned with the aggregate outcome of individual actions. Keynes's "consumption function," for example, which relates aggregate consumption to national income, is not built up from individual consumer behaviour; it is simply an empirical generalization. The focus is on income and expenditure flows rather than on markets. Purchasing power flows through the system from business investment to consumption, but it leaks out at two places in the form of personal and business savings. Counterbalancing the savings are investment expenditures in the form of new capital goods, houses, and so forth, which constitute a source of new injections of purchasing power in every period. Since savings and investments are carried out by different people for different motives, there is no reason why "leakages" and "injections" should be equal in every period. If they are not equal, national income, the sum of all income payments to the factors of production, will rise or fall in the next period. When planned savings equal planned investment, income will be at an equilibrium level, that is, a level at which it can sustain itself; when the plans of savers do not match those of investors, the level of income will go on changing until the two do match. One can complicate this simple model by making investment a function of the interest rate; by introducing the government budget, the money market, labour markets, imports and exports, foreign investment; and so forth. But all this is far removed from the problem of resource allocation and from the maximizing behaviour of individual economic agents.

The result is a kind of intellectual schizophrenia in which the techniques of microeconomics do not carry over fully into macroeconomics and vice versa. This is widely held to be an unsatisfactory state of affairs; economists have in recent years sought to build a bridge between the individual consumer and the overall consumption function and between the individual investor and the behaviour of aggregate investment. Nevertheless, the bridge remains incomplete, and the student of economics must be prepared to work with two boxes of tools.

FIELDS OF CONTEMPORARY ECONOMICS

The following is a bird's-eye view of the main fields of contemporary economics.

Public finance. Since the time of Ricardo economists have been concerned with the incidence of taxes, that is, with determining who it is that really pays a tax. If a corporation faced with a profits tax reacts by raising its prices, it may succeed in making the consumer pay the tax; on the other hand, if sales decline as a result of the rise in price the firm may have to lay off some of its workers, and the burden of the tax will be shared by consumers, wage earners, and shareholders. This simple example shows how complex may be the actual incidence of a tax. A large part of the literature of public finance in the 19th century was devoted to such problems.

Taxes and government expenditures

Keynesian economics brought new dimensions to public finance: the older preoccupation with tax incidence gave way to the analysis of the impact of government expenditures on the level of income and employment (see GOVERNMENT FINANCE). It was some time, however, before economists realized that they lacked a theory of government expenditures, that is, a set of criteria for determining what activities should be supported by governments and what should be the relative expenditure on each. One of the most exciting recent developments in the field of public finance is the attempt to devise such criteria. Decisions on public expenditures have proved to be susceptible to much of the traditional analysis of microeconomics. In the 1960s there developed a technique known as "cost–benefit analysis," which tries to appraise all of the economic costs and benefits, direct and indirect, of a particular activity so as to decide how to distribute a given public budget most effectively between different activities. This technique has been applied to everything from the construction of hydroelectric dams to the control of tuberculosis. Its exponents hope that the same type of analysis that has proved so fruitful in the past in analyzing individual choice may also succeed with problems of social choice.

Money. One of the oldest, widely accepted functions of government is control over the supply of money. The dramatic effects of changes in the quantity of money on the level of prices and the volume of economic activity were recognized and thoroughly analyzed in the 18th century, and monetary economics has ever since constituted one of the principal branches of economics. In the 19th century a complex and somewhat crudely formulated tradition grew up known as the "quantity theory of money," which held that any change in the supply of money can only be absorbed by variations in the general level of prices (the purchasing power of money). In consequence, prices will tend to change proportionately with the quantity of money in circulation. As the growth of fiat paper money gave governments increasingly effective control over the stock of circulating media, the quantity theory of money supplied an apparently simple rationale for the management of the economy: all that was needed to prevent inflation or deflation was to vary the quantity of money in circulation inversely with the level of prices.

Keynes's attack on monetary policy

One of the targets of Keynes's attack on traditional thinking in his *General Theory of Employment, Interest and Money* was this quantity theory of money. Keynes produced a different theory of the demand for money that implied that the impact of a change in the stock of money on the level of national income is weak and at best indirect; the effect on prices is virtually nil, he maintained, at least in economies with heavy unemployment such as prevailed in the 1930s. He put his emphasis instead on government budgetary and tax policy and direct control of investment. As a consequence, economists lost faith in monetary management and came to regard monetary policy as more or less ineffective in controlling the volume of economic activity.

In the 1960s there was a remarkable revival of the older view, at least among a small but growing school of American monetary economists. They accepted much of Keynesian economics but argued that the effects of fiscal policy are unreliable unless the quantity of money is regulated at the same time. They refurbished the quantity theory of money and tested the new version on a variety of data for different countries and for different time periods, leading to the broad conclusion that the quantity of money does matter (see MONEY).

In the late 20th century the controversy was still raging. It is notable that this debate, unlike previous debates in the history of monetary economics, was characterized by disputes over empirical findings—that is, it was focussed on the testable character of different monetary theories rather than on the manner of their formulation. Progress was made slower by the political overtones of the controversy: in some countries, belief in the efficacy of monetary policy had become a kind of litmus test of political conservatism. Nevertheless, a reconciliation between Keynesians and quantity theorists needed only some agreement as to the magnitude of monetary forces and the degree of stability of the demand for money. Monetary economics seemed at last to be coming of age as an empirical discipline.

International economics. The foundations of international economics were firmly established in the 19th century. The subject has consisted ever since of two distinct

but connected parts: (1) the "pure theory of international trade," which seeks to account for the gains obtained from trade and to explain how these gains are distributed among countries, and (2) the "theory of balance-of-payments adjustments," which analyzes the workings of the foreign exchange market, the effects of alterations in the exchange rate of a currency, and the relations between the balance of payments and the level of economic activity.

In modern times, the Ricardian pure theory of international trade has been reformulated by the American Paul Samuelson, improving on the earlier work of two Swedish economists, Eli Heckscher and Bertil Ohlin. The so-called Heckscher-Ohlin theory explains the pattern of international trade as determined by the relative land, labour, and capital endowments of countries: a country will tend to have a relative cost advantage in goods requiring the intensive use of the country's relatively abundant factor of production (thus land-rich Canada exports wheat) and to import goods requiring the intensive use of the country's relatively scarce factor (thus capital-poor Canada imports American automobiles). This theory absorbs Ricardo's law of comparative costs but goes beyond it in linking the pattern of trade to the economic structure of trading nations. It implies that foreign trade is a substitute for international movements of labour and capital—which raises the intriguing question of whether or not foreign trade may work to equalize the prices of all factors of production in all trading countries. Whatever the answer, the Heckscher-Ohlin theory provides a model for analyzing the effects of a change in trade on the industrial structures of economies and, in particular, on the distribution of income between factors of production. Much of the recent effort of specialists in international economics has gone toward refining the Heckscher-Ohlin model and testing it on an ever wider range of empirical evidence.

Labour. Like monetary and international economics, labour economics is an old economic speciality. It gets its raison d'être from the peculiarities of labour as a commodity. Labour itself is not bought and sold; rather, its services are hired and rented out. But since people cannot be disassociated from their services, various nonmonetary considerations play a role in the sale of labour services as contrasted with the sale of machine time or the rental of land. Yet, the bulk of the literature in labour economics was until recently concerned solely with the demand side of the labour market. Wages, the textbooks all said, were determined by the "marginal productivity of labour," that is, by the relationships of production and by consumer demand. If the supply of labour came into the picture at all, it was merely to allow for the presence of trade unions; unions could only raise wages by limiting the supply of labour.

The economics of labour and industry

After a long period of neglect, the supply side of the labour market began, in the 20th century, to attract the attention of economists. First, attention shifted from the individual worker to the household as a supplier of labour services; the increasing tendency of married women to enter the labour force and the wide disparities and fluctuations observed in the rate that females participate in a labour force drew attention to the fact that an individual's decision to supply labour is not independent of the size, age structure, and asset holdings of the household to which he or she belongs. Second, the new concept of "human capital"—that people make capital investments in their children and in themselves by incurring the costs of education and training, the costs of searching for better job opportunities, and the costs of migration to other labour markets—has served as a unifying explanation of the diverse activities of households in labour markets. In this way, capital theory (see ECONOMIC THEORY) has become the dominant analytical tool of the labour economists, replacing or supplementing the traditional theory of consumer behaviour. The economics of training and education, the economics of information, the economics of migration, the economics of health, and the economics of poverty are some of the by-products of this new perspective. A field that was at one time regarded as rather cut-and-dried has taken on new vitality.

Labour economics, old or new, has always regarded the explanation of wages as its principal task, including the factors determining the general level of wages in an economy and the reasons for wage differentials between industries and occupations. Wages are influenced by trade unions; the impact of their activities is of increased importance at a time when most governments manage the economy with one eye on the unemployment statistics. The prewar fears of chronic unemployment gave way to the postwar fears of chronic inflation at or near levels of full employment. In response to this a vast literature sprang up after 1945 analyzing the inflationary pressures stemming from both the supply side and the demand side of labour markets. Whether prices were being pushed up by the labour unions ("cost push") or pulled up by excess purchasing power ("demand pull") became the issues in this long debate on inflation, a controversy that is, of course, intimately related to the quarrels in monetary economics mentioned earlier.

Industrial organization. The principal concerns of this field are the structure of markets, public policy toward monopoly, the regulation of public utilities, and, of late, the economics of technical change.

The monopoly problem, or, more precisely, the problem of the maintenance of competition, does not fit well into the received body of economic thought. Economics started out, after all, as the theory of competitive enterprise, and even today its most impressive theorems require the assumption of numerous small firms, each having a negligible influence on price. Yet the typical market structure of manufacturing today is that of oligopoly—competition among the few—and some industries are dominated by firms so large that their annual sales volume exceeds the national income of the smaller countries of western Europe. It is tempting to leap to the conclusion that oligopoly is deleterious to economic welfare, on the ground that it leads to the misallocation of resources. But some economists, notably Joseph Schumpeter, have argued that economic growth and technical progress are brought about not through free competition but through large firms and the destruction of competition. According to this view, monopoly has its origin in the need of business firms to protect themselves from the risks associated with the introduction of new products, new techniques, and new methods of marketing. The giants, therefore, compete not in price but in successful innovation, and this kind of competition has proved more effective for economic progress than the more traditional price competition described in orthodox textbooks of economic theory.

Merits of monopoly and competition

Although this thesis smacks of *post hoc ergo propter hoc* ("after this, therefore because of this")—giant firms have prospered in rapidly growing economies; therefore, growth is due to giant firms—it makes the merits of "trust busting" and cartel dissolution somewhat less compelling. The question is what sort of competition is socially most desirable. If each of four or five large firms in an oligopolistic industry finds it necessary to compete in terms of the quality of its products and its research or by means of better technology and superior merchandising, the performance of the industry may well be more satisfactory than if it were reorganized into a price-competitive industry. But if the four or five giants settle down to a quiet life and concentrate their rivalry on sales promotion techniques, the verdict must be less favourable. One cannot, it seems, draw facile conclusions about the competitive results of different market structures; it is necessary to approach the monopoly problem with a generous dose of pragmatism.

The reason why there is so much uncertainty in the economic discussion of policies toward big business is the lack of a general theory of oligopoly. There are dozens of special theories applying to special cases, but there is no single, organizing framework with testable implications about the behaviour of oligopolists in general.

Agriculture. Farming has long provided economists with their favourite example of a perfectly competitive industry. But with increasing government regulation of agriculture, it also provides striking examples of the effects of price controls, income supports, output ceilings, and marketing cartels. Agricultural economics commands attention wherever governments wish to stimulate farming or to protect farmers—which is to say, everywhere.

Agricultural economists generally have been closer to their subject matter than other economists. In consequence, more is known about the technology of agriculture, the nature of farming costs, and the demand for agricultural goods than is known about any other industry. The student of economics who wants to learn how to estimate a production function or a demand curve is well advised to go to the literature on agricultural economics.

The underdeveloped countries have furnished a new laboratory for agricultural economics. Many such countries have a "subsistence agriculture," in which the farmer produces mainly for his own family consumption and brings to market only what is left over. Subsistence farmers are only tenuously linked to the money economy. They are more reluctant than commercial farmers to take the risks entailed in experimenting with new seeds, fertilizers, and farming methods: given the vagaries of the weather, they prefer to operate on the basis of the worst weather that can be expected rather than the best or even the average. While the pessimism of subsistence farmers is perfectly rational, it makes the task of predicting their response to new prices or new methods doubly difficult. The economist must also recognize that every agricultural area presents its own production, processing, and marketing problems. To suggest changes that will raise agricultural productivity in these circumstances is no easy task, and agricultural economists often find themselves warning governments not to intervene too vigorously, since much intervention in the past has been shown to be wrong or inappropriate.

The problem of economic development in Africa, Asia, and Latin America centres on the agricultural sector; one of the abiding questions is how far industrialization can proceed without there first being an agricultural revolution. For this reason, if for no other, agricultural economics has a large future.

Growth and development. The study of economic growth and development is not a single branch of economics but falls, in fact, into two quite different fields. The two fields—"growth" and "development"—employ different methods of analysis and are indeed addressed to two distinct types of inquiry.

Development economics is easy to describe. It is one of the three major subfields of economics, the other two being microeconomics and macroeconomics. Development economics resembles economic history in that it seeks to explain the changes that take place in economic systems with the passage of time.

The subject of economic growth is not so easy to characterize. It is the most technically demanding field in the whole of modern economics, impossible to grasp for anyone who lacks differential calculus. Its focus is the properties of equilibrium paths, rather than equilibrium states. One makes a model of the economy and puts it into motion, requiring that the time paths described by the variables be self-sustaining in the sense that they continue to be related to each other in certain characteristic ways. Then one can investigate the way economics might approach and reach these steady-state growth paths from given starting points. Beautiful and frequently surprising theorems have emerged from this experience, but as yet there are no really testable implications nor even definite insights into how economies grow.

Harrod–Domar growth model

Growth theory began with the work of Roy Harrod in England and Evsey Domar in the United States. Their joint product has been known ever since as the Harrod–Domar model. Keynes had shown that new investment has a multiplicative effect on income and that the increased income generates extra savings to match the extra investment, without which the higher income level could not be sustained. One may think of this as being repeated from period to period, remembering that investment, apart from raising income disproportionately, also generates the capacity to produce more output that cannot be sold unless there is more demand, that is, more consumption and more investment. That is all there is to the model. It contains one behavioral condition—that people tend to save a certain proportion of extra income, a tendency that can be measured. It contains one technical condition—that investment generates additional output, a fact that can be established. And it contains one equilibrium condition—that planned saving must equal planned investment in every period if the income level of the period is to be sustained. Given these three conditions, the model generates a time path of income and even indicates what will happen if income falls off the path.

More complex models have since been built, incorporating different saving ratios for different groups in the population, technical conditions for each industry, definite assumptions about the character of technical progress in the economy, monetary and financial equations, and much more.

Mathematical economics. Differential calculus has long been the traditional tool of mathematical economics. Many economic problems, particularly in microeconomics, take the form of maximizing some variable (such as profits) subject to a constraint (such as the production function), for which calculus supplies the simplest technique. Traditionally it was applied to problems in comparative statics. These problems include so-called endogenous variables, the values of which are determined within the model, as well as constants that originate outside the model and are called "exogenous variables" or "parameters." The object is to discover the effects of changes in one or more of the parameters upon the equilibrium situation. The latter is a situation in which all of the endogenous variables are simultaneously in a state of rest. If the value of some of the parameters is changed, the result is a new equilibrium state.

The application of mathematical methods

Much economic analysis, even when it is expressed in words, is simply the method of comparative statics. But comparative statics has its limitations: it tells the investigator *where* the system will arrive, but it does not tell him *when* it will arrive or what will happen along the way; and it cannot tell him whether, once driven out of the way, it will ever get back to its destination. In other words, comparative statics ignores the process of adjustment from the old equilibrium state to the new one, and it entirely neglects the time element in that adjustment process. The study of this process of adjustment over time is called "economic dynamics," and one may think of it as the economics of disequilibria.

Just as differential calculus is the mathematics of comparative statics, difference and differential equations are the ideal tools for handling dynamic problems. Difference equations deal with time as a discrete variable—changing only from period to period—whereas differential equations treat time as a continuous variable; the choice between them is simply one of convenience. They enable one to ask such questions as: if the system is pushed out of equilibrium, perhaps because one of the parameters of the model has changed, will economic forces drive it toward a new equilibrium position or away from one, will the time path described by the endogenous variables be steady or fluctuating, and if fluctuating, will the movements be damped down or will they increase and become explosive?

Economic dynamics is one of the newer developments of mathematical economics, and often it falls short of the ambitious demands made on it. Dynamic models, for example, are typically formulated in terms of linear equations, not because the world is linear but because nonlinear equations can be very difficult to solve. Likewise, the coefficients of difference and differential equations are usually taken to be constants, again for the sake of making the mathematics of the analysis manageable. This means that if the economic environment changes as the model runs its course, its predictions will be false. An abiding danger in all mathematical economics is the tendency to adopt economic assumptions for the sake of mathematical convenience. The way to meet this danger is for economists to acquire enough mathematical sophistication so that they will not be dazzled by displays of mathematical technique.

Econometrics. Like mathematical economics, econometrics is something economists do rather than a special area of interest. Econometrics refers to the study of empirical data by statistical methods, the purpose of which is the testing of hypotheses and the estimation of relationships suggested by economic theory. Whereas mathematical economics considers the purely theoretical aspects of

economic analysis, econometrics attempts to falsify theories that are expressed in explicit mathematical terms. But frequently the two go together.

The classic technique for estimating an economic relationship is that of "least squares," which is a method of fitting a trend line to a scatter of observations that minimizes the square of the deviations of the observed points from the line. To take a simple example: the Keynesian theory assumes that consumers' expenditures depend principally on income; one may interpret this to mean that consumption depends *only* on income and then test the hypothesis by trying to fit a trend line to a series of observations of income and consumption over a period of time. In so doing, one is really saying that the observations that fall to either side of the line are due either to errors in measuring the variables or to errors in specifying the relationship between consumption and income. It is essential to the method of least squares that these "errors" be randomly distributed or at any rate distributed in known ways. When this condition is violated, least squares estimates are unreliable. It is sometimes difficult to tell with economic data just how the errors are randomly distributed, and it is precisely for this reason that an econometrician is needed rather than an ordinary statistician.

Problem of error distribution

A still more significant trend in recent econometrics is the tendency to move from single-equation estimates (such as the relationship between consumption and income) to systems of simultaneous equations. While consumption depends on income, income also depends on consumption; this kind of interdependence requires two equations rather than one. More generally, most economic variables are the result of demand and supply forces that simultaneously determine quantities and prices. To estimate a demand curve for butter from a single-equation regression (by relating the price of butter to the quantities of butter consumed, the incomes of consumers, and the prices of near substitutes for butter) is likely to produce a biassed answer because the price of butter is also influenced by supply conditions in the dairy industry. This creates the so-called identification problem, namely, the question of whether it is possible to identify a demand curve or a supply curve from observed price–quantity data. The use of simultaneous equation models to estimate economic relationships is by now perhaps the best way of distinguishing econometrics from economic statistics.

The foregoing discussion covers only nine major branches of economics. There are many other fields in economics, including economic history, comparative economic systems, business cycles, economic forecasting, national income accounting, managerial economics, business finance, marketing, the economics of natural resources, economic geography, consumer economics, and regional economics.

(M.Bg.)

Political science

Political science is most generally understood to mean the systematic study of government processes by the application of scientific methods of analysis. More narrowly and more traditionally, it has been thought of as the study of the state and of the organs and institutions through which the state functions. In most countries, political science is thought to be a single discipline, but the plural form has been used in France, as in the name of the École Libre des Sciences Politiques (now Institut d'Études Politiques de l'Université de Paris), founded in 1871—although there is also an Association Française de Science Politique. Speculation about political subjects is not unknown in ancient non-Western cultures, but most students agree that the roots of political science are to be found in the earliest sources of Western thought, especially in the works of Aristotle, who is recognized by many as the founder of political science.

Political science and political philosophy

Although political science may be distinguished from political philosophy, the distinctions are unsatisfactory inasmuch as they lack categorical rigour. In the most usual distinction, political philosophy is thought to be concerned primarily with the study of political ideas, often within the context of their times. It is strongly normative in its thrust and disposition and rationalistic in its method. Political science, however, concerns itself with institutions and behaviour, eschews normative judgments as much as possible, and attempts to derive principles from objective facts with as much quantification as the evidence will allow. Political philosophy thus speculates about the place and order of values, the principles of political obligation (why men should or should not obey political authority), and the nature of such terms as right, justice, and freedom. Political science, on the other hand, seeks to establish by observation (and, if possible, by measurement) the existence of uniformities in political behaviour and to draw correct inferences from these data. The stated differences between political philosophy and political science are less than is sometimes supposed, for the most empirical scholar in both the social and natural sciences makes use of unproved postulates, hunches, and intuitions; and the most rationalistic philosopher employs conceptions that embody empirical statements.

Some theorists, however, who believe it possible to develop a completely value-free science of politics insist that the distinction between political philosophy and political science is not faint but vivid. It is their opinion that only a few hundred philosophers and political theorists have ever contributed to systematic speculation about and study of politics. It was said in 1966 by one of the exponents of this view that "probably two out of every three political scientists who have ever lived are alive and practicing today." In this view, political philosophy is believed to be addicted to obscurity and opacity of statement—an effort to bespeak the unspeakable—and it is thought that its task should be a more modest one, namely, to grope with the grammar of philosophical statements. This would require political philosophers to put their intelligence to the elucidation of the language of politics and to expose the difficulties placed by language in the consideration of matters of fact. One of those interested in this approach was the English political philosopher T.D. Weldon in *The Vocabulary of Politics* (1953), but few have followed his initiative.

Scientific status of the subject

The question as to whether political science is a science is largely inconsequential, because the problem is primarily one of definition. If the term science is to be applied to any body of systematically organized knowledge based on facts ascertained by empirical methods and described by as much measurement as the material allows, then political science is a science, just as are the other social disciplines. If, on the other hand, the term science is to be limited to those disciplines in which the scholar can control the materials to be studied and can perform experiments that others can reproduce under the same conditions and in which predictability is possible, then the label is less appropriate, although not entirely misapplied. The American political economist Thorstein Veblen denied that political science was anything more than a "taxonomy of credenda," and, more recently, a British writer, Bernard Crick, has said that the hope of creating an artificial science of politics on natural principles, although not originally and uniquely American, has been largely generated and sustained by two aspects of the American culture—an agreement on liberal doctrine that has made politics less a matter of serious doctrinal splits than of mere disagreement among partisans of the same creed, and a general national preoccupation with technology.

HISTORICAL DEVELOPMENT OF POLITICAL SCIENCE

Early trends. The origins of contemporary political science are to be found in the enthusiasm for the creation of social science that was widespread in the 19th century, an enthusiasm stimulated by the rapid growth of the natural sciences. It might be said that one starting point for the development of modern political science is the work of the Comte Henri de Saint-Simon, a notable Utopian Socialist, who in 1813 suggested that morals and politics could become "positive" sciences; that is, disciplines whose authority to command belief would rest not upon subjective preconceptions but upon objective evidence. With him worked the mathematician and philosopher Auguste Comte, the two collaborating in the publication in 1822 of the *Plan of the Scientific Operations Necessary for the*

Reorganization of Society, which argued, among much else, that politics would become social physics and that the purpose of social physics was to discover unchanging laws of progress. Out of this collaboration emerged the law of the three stages through which knowledge had to pass—the theological, the metaphysical, and the positive—that Comte was to establish as the theme of the science of social physics, a study he came to name sociology. An intellectual connection between political science and sociology was thus early established in schemes of political and social regeneration and reform, although political science was thought to be limited to only one form of association in society, namely, the state. Comte thought that the principal methods for the study of social phenomena were observation, experiment, and abstraction. Although one might have thought that politics could not be an experimental science, Comte was of the view that political experimentation did take place whenever there was a change in the life of the state, intended or not. It must be said, however, that even on this account there is no close similarity to experimentation in chemistry and physics, in which all the variables can be controlled.

In the search for more objective methods of inquiry into political and other social phenomena in the 19th century, contributions to the explanation of the state were supplied by several new intellectual disciplines. Because political science deals with some aspects of human behaviour, for example, it is closely allied to other social sciences that also deal with human behaviour. Long before the development of scientific inquiries in the 19th century, numerous theories of the state had drawn inspiration from the human being as model, as in the *Policraticus* of John of Salisbury (1159), in which the physiology of the body and that of the state are compared; or in *The Republic* of Plato, in which the elements of the human personality prefigure the class structure of the state; and in Rousseau's *Contrat Social,* in which the political order is animated by a general will (will being a human attribute). The positivism of the 19th century, however, brought new approaches to the study of the state, although the older ones continued to coexist with them. Among those following Comte was the Polish-born sociologist Ludwig Gumplowicz, who built a sociology on Comtean foundations but who owed much also to Darwin, to the social Darwinist Herbert Spencer, and to contemporary anthropology. In Gumplowicz' view, the earliest forms of group life were small hordes bound by consanguinity, which developed into matriarchies and patriarchies. He supposed the existence of a social-evolutionary process characterized by conflicts between autonomous groups and by conflicts of interest within those groups. The product of this process was the state, founded on force and maintained by power.

Sociological analyses of the state

Gumplowicz and several other 19th-century political sociologists anticipated 20th-century concerns in political science with the significance of groups, the nature of interests, the role of parties as interest groups, and the social context within which political events occur.

Still another 19th-century writer with some precedental connection to the political science of the 20th was the Italian Vilfredo Pareto. Although he lived and wrote in both the 19th and 20th centuries, he may be counted in the earlier period because of his advocacy of the "logico-experimental" approach to sociology, which involved observation and logical inference. Pareto had no direct influence on the development of political science, but in two respects his sociology had implications for the developing discipline. First, his was a psychological sociology, and much of his concern was with the influence of beliefs, attitudes, opinions, and sentiments in shaping social life. This anticipates the 20th-century approach of many political scientists who regard psychology as the most important adjunct of the scientific study of politics. Second, Pareto thought of society as a system always tending toward equilibrium, and the conception of politics as a system was to mark much of academic political science after World War II (see below). Also to be mentioned in the context of important 19th-century sociological theories is the work of a Swedish political scientist, Rudolf Kjellen, whose systematic treatment of the state as a fusion of organic and intellectual moral elements in an ambience of geographical determinism led to a theory of politics that he termed geopolitics.

Juristic influences. In addition to sociology, another discipline with which the development of political science has been historically related is law. A close connection between the state and law was made in the 16th century in the French political philosopher Jean Bodin's theory of sovereignty, which supposed the necessary existence in each state of an authority to make the law. This theory gave rise to numerous juristic theories of the state, especially among German publicists of the 19th century who sought to define the nature of federation and empire. Although the doctrine of sovereignty made most sense as a statement of the power of monarchs to make the law in simple unitary (nonfederal) systems, earnest writers forced the facts of federation and empire to fit the theory and often ignored facts that did not. The Bodinian view invested sovereignty with attributes of singularity and omnipotence—that is, sovereignty was viewed as being indivisible and absolute—but theoretical efforts were made in Germany before 1871, notably by Georg Waitz, to establish federal theory on a division of sovereignty between the centre and the member states. After 1871, the theory of Max von Seydel, namely, that since sovereignty is indivisible, it cannot be divided and must rest in a single location—either in the constituent members of a federal system or in the centre—supplanted that of Waitz. Georg Jellinek, an Austrian, attempted to resolve the paradox that so-called sovereign states are in fact limited by constitutions and laws and by membership in the family of nations by arguing that the restrictions are autolimitations. That is to say, since by definition there is no power above the state, if in fact there are limitations it must be the state that created them. The state therefore continues to be omnipotent. In France, juristic theories of the state also persisted strongly because the teaching of political science was conducted in the law schools as "constitutional law." Efforts were made by some professors, however, to broaden the legal approach, notably by Léon Duguit, who attempted a sociological positivistic treatment of juridical rules for the limitation of the state, and Maurice Hauriou, who contributed a theory of institutions. In England the emergence of political science as a subject was recognized in the establishment of the London School of Economics and Political Science in 1895 and by the founding of a separate chair of politics at Oxford in 1912.

The doctrine of sovereignty

Developments in the United States. The enthusiastic development of social sciences in the 19th century, stimulated as it had been by the rapid growth of the natural sciences, reinforced an existing interest in politics in the United States and created a generation of distinguished American political scientists. There had, in fact, been much interest in the teaching of political subjects in American colleges and universities well before the 19th century. Political science in the United States, however, free of any connection with moral philosophy, fusion with history, or submergence in political economy, may be said to date from 1880, when John W. Burgess, after studying at the École Libre des Sciences Politiques, succeeded in establishing a separate school of political science at Columbia University. Although political-science faculties increased in numbers after 1900, the growth was uneven, and in some major institutions separate departments were not created until after World War I.

The development of American political science in the last quarter of the 19th century was influenced by the experience of numerous scholars who had done graduate work at German universities in which political science was taught as *Staatswissenschaft* ("science of the state") in an ordered, structured, and analytical organization of concepts, definitions, comparisons, and inferences. To modern readers the work of these men often seems somewhat formalistic and institutional in tone and focus. It did represent, however, an effort to establish an autonomous discipline, separate from history, moral philosophy, and political economy. Some of the new American political scientists, such as Woodrow Wilson and Frank Goodnow, also showed in their writing an awareness of such new

intellectual currents as theories of evolution and turned their attention to an examination of American institutions that, in the United States Constitution, had originally been based on the admiration of much of the 18th-century world for the harmonious perfections of mechanics. For Wilson, it was to be Darwin, not Newton, who would provide the inspiration for a transformation in American political science from the study of static institutions to the study of social facts, more truly in the positivist temper, less in the analytic tradition, and more oriented toward factual realism.

Bentley's early work on group processes

Little-noticed at the time he wrote *The Process of Government* (1908), Arthur F. Bentley was to have a celebrated influence on the development of American political science in the 1930s and the 1950s. Although his influence was not immediately felt, he sounded certain themes that became the orthodoxy of the new science of politics after World War II. First, he rejected all metaphysics and normative formulations as "spooks" and "mind-stuff" and insisted that the proper study of politics was observable fact, in imitation of the natural sciences. Second, his basic concept was the "group," rejecting thereby all previous formulations of the subject matter of political science that centred on the "state." Third, the "raw material of government," he thought, was the activity of men and the processes through which this activity flowed in legislation, administration, and adjudication. Behaviour and process were to become the focus of much of the interest in political science in the 1950s, after Bentley's single work on politics was given new currency just before the outbreak of World War II. He was considerably before his time in presaging the end of preoccupation with institutional and descriptive analysis. He acknowledged the leadership of Gumplowicz, who, in his view, had "taken the most important step toward bringing out clearly the nature of the group process," discarding the "individual as a causal factor in society," and insisting "that all social movements are brought about by group interaction."

Although the effort of Bentley to develop an objective, value-free analysis of politics had no initial consequence, other movements toward this goal were more immediately successful. The principal impetus was provided by what became known as the Chicago school in the mid-1920s and thereafter. The leading figure in this movement was Charles E. Merriam, who in 1925 published *New Aspects of Politics,* a book that argued for a reconstruction of method in political analysis, urged the greater use of statistics in the aid of empirical observation and measurement, and postulated that out of the converging interests of politics, medicine, psychiatry, and psychology might come "intelligent social control." The basic political datum for Merriam at this stage of his thinking was "attitude"; hence his reliance upon the insights of psychology for a better understanding of politics. These ideas were not entirely new, since Graham Wallas, an Englishman, had said in *Human Nature in Politics* (1908) that a new political science should be based upon quantitative methods and that serious attention should be given to the psychological elements ("human nature") in political activity, including nonrational acts and the exploitation in political life of subconscious nonrational inferences. The American political scientist and journalist Walter Lippmann had expressed much the same view in *Public Opinion* (1922). One of those in the Chicago group who carried the connection between politics and psychology quite far was Harold Lasswell, in his *Psychopathology and Politics* (1930). In *Power and Personality* (1948) he fused the Freudian categories of the earlier work with subsequent writings on power.

Influential works of Merriam and Lasswell

These two leading expositors of the Chicago school, Merriam and Lasswell, both published books at about the same time that gave a central place to the phenomenon of power in the empirical study of politics. Merriam published *Political Power* in 1934 and Lasswell *Politics: Who Gets What, When, How* in 1936. Merriam undertook to show how power came into being, to describe what he called the credenda, miranda, and agenda of authority (which he tended to equate with power), the techniques of power holders, the defense available to those over whom power is wielded, and the dissipation of power. Lasswell's 1936 work was a naturalistic description of "influence and the influential." Although both were cast in the empirical mode, the second was more successful in this regard than the first, which tended to be abstract and rhetorical. A truly empirical work of the Chicago school that had considerable significance in the development of academic political science was Charles E. Merriam and Harold F. Gosnell's work, published in 1924, on *Non-Voting, Causes and Methods of Control,* which used sampling methods and survey data. Since then, certainly one of the most successful achievements in empirical political science has been the study of voter behaviour and election results. Although members of the Chicago school insistently professed an interest in value-free political science, they were characterized by two normative predilections—their acceptance of the values of the democratic system and their earnest attempts to improve it through their writings.

By 1945 political science in the United States was much more than the concern for institutions, law, formal structures of public government, procedures, and rules that it had earlier been. There was by then also a considerable body of writing on processes—on the dynamics as well as the statics of public governance. There were works on pressure groups and lobbies, on the "invisible government" operating behind public authority, on actual bureaucratic processes as distinguished from the rules of administrative procedure, on bosses and political parties, and on ethnic influences in the behaviour of the electorate. All of these works signified a turning away from formality and the development of a growing interest in the factual realities of political behaviour.

Developments in Europe. Outside the United States the development of political science was less quantitative and behavioral. In France the main tradition in the study of political institutions was still legal rather than sociological. Constitutional law and history were abundant, and politics in the main was viewed from legal perspectives, though not entirely so. As early as 1913 one writer, André Siegfried, had introduced the geographical and historical study of elections in *Tableau politique de la France de l'Ouest sous la troisième République,* and the development of French sociology in the works of Émile Durkheim and others had applications in political science.

In England, although there was little development of a quantitative science of politics, a substantial contribution was made toward the formulation of political philosophies that centred attention upon the group basis of politics. Particularly notable were the works of J.N. Figgis, G.D.H. Cole, Ernest Barker, and Harold Laski. In Sweden, Herbert Tingsten in his work *Political Behaviour: Studies in Election Statistics* (1937) gave currency in the title to what was to be the main development in political science after World War II.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY POLITICAL SCIENCE

Behavioralism. In American political science since the end of World War II, the behavioral persuasion has been the dominant one. A former president of the American Political Science Association has attributed the rapid development of the behavioral approach to six causes: the inspiration of the Chicago school; the immigration to the United States in the 1930s of large numbers of European scholars (particularly Germans) with backgrounds in European sociology, who stressed the relevance of sociology to politics; the movement of many political scientists into administrative and political positions during World War II; the influence of foundation support in the encouragement of research in political behaviour; the increasing development of the survey method in certain political studies, such as voter behaviour; and the missionary work of the Social Science Research Council under leadership sympathetic to behavioralism.

Etiology of the behavioral approach

Although the term behavioralism has been freely used in political-science writings, there is in fact confusion as to whether it is a field of study, a method, or an approach. One American political scientist, Heinz Eulau, in *The Behavioral Persuasion in Politics* (1963), has said that the behavioral persuasion "is concerned with what man does

politically and the meanings he attaches to his behavior," and he has suggested that researchers cannot afford to get tangled up in problems of definition. Another American, Robert Dahl, has said that it is a mood or even "the scientific outlook." The term behavioral, then, may be merely a term having distinctiveness, weight, and value for a certain time only, since it seems primarily to signify that phase in the quarter century after World War II during which there was a significant revival of interest in empirical studies in politics, a movement strong enough to establish at least a partnership with the traditional approaches, although some of its advocates have gone so far as to say that their science has made traditional approaches outdated.

Systems analysis. Contemporary political science does not display as much coherence as earlier modes of inquiry (coherence was one of their principal virtues), but it does exhibit a refreshing complexity, of both concept and method. Although much effort has been spent on the production of empirically derived evidence in a scatter of fields and a diffusion of subjects and although critics have scoffed at studies that they have said only prove the obvious, the product of 25 years of behavioralism has been voluminous and often stimulating in quality. Common in this output is the assumption that politics is process, the ceaseless interaction of individuals and groups on each other in a flow of activities in and around public governance. The most commanding concept devised to fit this flow with form has been the concept of the system. In this view the focus of political science is not the individual in solitude nor the multitude nor even the group. The American political scientist David Easton's *The Political System* (1953), a seminal work in empirical political theory, conceived of the political system as part of the total social system. For Easton the political system comprised all those activities having to do with the formulation and execution of social policy; that is, "policy-making process." The identifying criterion of political behaviour for Easton was the "authoritative allocation of values" for the society, by which he distinguished his sense of the subject matter of political science from that of Lasswell, who had argued that political science was concerned with the distribution and contet of patterns of value throughout the society. In this conception of system, which is not unrelated to the conception of system in physics and biology, linkages are found between the system and its environment. Inputs (demands) flow into the system and are converted into the outputs (decisions and actions) that constitute the authoritative allocation of values for the society; that is, the distribution of rewards in wealth, power, and status that the system may provide.

Variants of systems theory

There are various versions of systems theory, although all suppose the existence of the system and more or less consciously pattern it after systems models in other disciplines. There has thus been developed a view of systems, based upon engineering models and ideas, that rejects the view of Easton that there can be a general theory of political systems. Another theorist, drawing on cybernetics, has viewed the political system as a communications net. Although expositors of systems concepts generally suppose they are working with scientific theory, critics have said that they are merely creating new taxonomies. What is certain is that systems theory has introduced new and—to more traditional scholars in the field—unusual terms into the vocabulary of politics. Thus, instead of such terms as the state and sovereignty, the new idiom speaks of systems, inputs and outputs, feedbacks, circular loops, netwoks, legitimacy symbols, information storage and retrieval, political socialization, interest articulation and aggregation, cluster blocs, zero-sum games, macropolitics and micropolitics, and much else drawn from the concepts and languages of other sciences and from statistics, with hopeful assertions about the possibility of making politics predictable.

Although decision making as a research field in political science may be analytically regarded as part of the process by which inputs are converted into outputs in systems theory, it also has an origin independent of systems theory. Interest in game theory provided a stimulant to studies of decision making, which already had an established place in the lore and lexicon of politics and administration. In decision theory the paradigm is the rational actor, and the research problem is the ascertainment of the most efficient means to effect given goals, which may or may not be rational. Investigators into decision making are generally less concerned with the system as a whole than with the activity that they regard as the most important part of it, namely, the way in which decisions are or should be made according to rational calculation. A body of writing on decision making by judges has earned a sub-classification as "jurimetrics."

Interest groups, elites, and political parties. Studies of interest groups, elites, and political parties also have their own independent origins, although they, too, have been brought within the framework of systems analysis. Interest groups and political parties, for example, have been described as agencies for the articulation and aggregation of interests, which in turn provide inputs (demands) for the political system to convert into outputs (decisions and actions). But interest-group analysis antedates the advent of behavioralism. The modern concern for the subject starts perhaps with studies of prohibition and other pressure groups during the 1920s. More generalized and theoretical treatments of interest groups and political parties resulted in part from a revived interest in the work of Bentley in the 1930s and 1950s. The study of elites was begun at least as early as 1936 (Lasswell's "influentials"), and it came to the forefront in the 1950s in community studies made by political scientists in Atlanta, Chicago, New York, New Haven, and elsewhere. Such studies had long been familiar in sociology, but they acquired special significance in political science, because democratic values seemed threatened by the possible existence of elites. Despite declarations of their concern to establish a value-free science, American political scientists have thus tended to cleave to the traditional democratic ethic, in which elites are presumed to have no place.

Analysis of political attitudes and voting behaviour. What some have called the behavioral revolution had its greatest successes in the analysis of public opinion, political attitudes, and electoral behaviour. Especially in the period after World War II, the refinement of statistical techniques in public-opinion polling, the analysis of voter behaviour, and the development of new research concepts have brought the study of opinions and attitudes closest to the goal of the scientific outlook and some considerable distance from the mark made by Merriam in the 1920s. The Survey Research Center at the University of Michigan has become an important national centre for the collection of data on elections and voter behaviour.

CURRENT TRENDS

The effort to establish a value-free, objective political science in the United States in the two decades after World War II has won what is doubtless a permanent place for the scientific study of politics, but it has also bred a critical reaction. In the late 1960s opponents of "scientism" rejected what they felt was the increasing subjection of spontaneity and human values to determinisms in every aspect of life and argued that political science was an example of the pervasiveness of technology and of a search for rationality in a social complex that might be irrational and out of control. Although political science had developed skillful and sensitive techniques for the quantification of data termed political, there is no orthodoxy on the scope of political science nor on the delimitation of separate areas of research. Quite different outcomes emerge from the basic assumptions as to the focus of the discipline, whether it be power, government, system, process, decision making, or policy formulation. The failure of the discipline to settle the question of identity has made for a creative and inventive exploration of many avenues of research, but the achievement of a unified general theory of political behaviour on which there is common consent still lies ahead.

Reactions against "scientism"

Even outside the United States, non-normative political science is not extensive, although it is not unknown. England has made many creative contributions to political theory and law throughout its history, as have other Euro-

pean countries, but it also has produced substantial works that can be classified as positive (*i.e.,* nonnormative) political science. Among these are the studies of R.T. McKenzie and D.E. Butler in the field of political parties, voter behaviour, and pressure groups, S.E. Finer on interest groups, and U.W. Kitzinger on German elections. Notable work has been done in France in the field of political parties by Maurice Duverger and François Goguel. Work also has been done on the political socialization of children in French schools by Charles Roig and François Billon-Grand. In Denmark beginnings were made in the systematic study of political science in 1959 with the founding of the Institute of Governmental Studies at the University of Århus. In Finland the first extensive studies on voting behaviour appeared in 1956.

In Japan there was a flowering of social sciences after World War II, but political science at first did not grow at the same pace as other social sciences because of lack of agreement about both subject matter and method, a difficulty felt from the beginning by the man recognized to be the founder of Japanese academic political science, Onozuka Kiheiji, who published *Principles of Political Science* in 1903. With the opening of the behaviour of public officials and institutions to scientific scrutiny, however, there has been much political inquiry that would qualify as political science, including, for example, systems analysis and works on political culture, political development, and process and behaviour.

Political science in Communist countries

Although in the past the objective study of political subjects by researchers in Communist regimes has been difficult, if not impossible, a somewhat more permissive policy in some countries has led to what may be the beginnings of scholarly political science. The most advanced political science is to be found in Poland and Yugoslavia, and Romania, the Czech Republic, and Slovakia have come to recognize political science as a discipline. The former Soviet Union did not sanction political science, but scholars did conduct empirical research, which was endorsed in 1962 by the U.S.S.R. Academy of Sciences under the term "concrete sociological investigations." The greatest movement has been in the conduct of public-opinion polls using advanced Western techniques. Interest in the development of political science was evidenced in the publication in 1969 of *Politicheskaya nauka v SShA: Kritika burzhuaznykh Kontseptsy vlasti* ("Political Science in the U.S.A.: A Critique of Bourgeois Conceptions of Power"), by V.G. Kalensky. Although there was no officially sanctioned political science, there was a Soviet Association of Political Sciences, which sent delegates to the meeting of the International Political Science Association. Despite its title, the Soviet Association of Political Sciences was most heavily oriented toward the state and the law, and its members were critical of what they regarded as the anti-Marxist bias of bourgeois political science. One of its members, however, F.M. Burlatsky, published a major article in *Pravda* in 1965 calling for the establishment of a genuine political science in which the findings would emerge from the data.

In Yugoslavia a political-science association was established in 1951, and in 1962 a faculty of the political sciences was established at the University of Zagreb. In Czechoslovakia a political-science association was formed in 1965 and became a member of the International Political Science Association; and in 1968 a political-science association was formed in Romania. Political science in Yugoslav universities has tended to centre on traditional divisions of the discipline, such as political theory, comparative government, and international relations. In Poland political science has centred on the study of political behaviour, on community power structures, on voter behaviour, and on public opinion. Techniques of considerable sophistication have been employed.

Political science is one of the means by which people seek to understand the human condition and man's fate. Or, as Aristotle believed, politics is the most important of human activities, and the sovereign science. For 24 centuries, at least, the greatest intellects and scholars have striven to state the universal elements of just order in human affairs. None has succeeded in this, so far, hopeless ambition, although most have contributed some special insight and added to the common wisdom. The effort to achieve magisterial comprehension will doubtless continue; and the search will change direction as experience requires, with the aid that new perceptions, concepts, and methods will provide. Progress toward establishing general laws, however, may never be as steady or as swift in political science as it has been in laboratory sciences like physics; for, as Albert Einstein once said, politics is more difficult than physics. (Ea.L.)

Study of international relations

HISTORICAL DEVELOPMENT OF THE STUDY OF INTERNATIONAL RELATIONS

Three contemporary forces go far to account for the impressive growth of scholarly studies of international relations and foreign policy in the 20th century. An autonomous academic discipline has emerged, related to geography, history, law, sociology, psychology, general political science, philosophy, and other fields, yet belonging to no one of these. The first impelling force was noted above: the growing demand to find better, less dangerous, more effective means of guiding relations among peoples, societies, governments, and economies. A second force is the result of the monumental upwelling of intellectual activity in modern times based on the belief that systematic observation and inquiry will dispel ignorance and serve the betterment of mankind. One sentence by Diderot characterizes this force as well today as when it was written in the 18th century: "Everything must be examined, everything must be shaken up, without exception and without circumspection." The third force is the consequence of the popularization of political affairs, including one of its most important sectors, foreign affairs.

The popularization of foreign and military affairs

Only late in the 19th century did the traditional view that foreign and military matters should remain an exclusive preserve of rulers and special elites yield to the opposite belief that such matters constitute an important concern and responsibility of all the people. This popularization of international relations made logical the idea that education should include instruction in foreign affairs and that knowledge in the field should be advanced in the interests of public control over international political and military matters. The experience of World War I, the war to make the world safe for democracy, strengthened the conviction that not enough was known about international relations and that the universities should reduce ignorance in the field through more research and teaching.

Between the two world wars. A strong impulse toward the development of international studies in universities came in the 1920s. New centres, institutes, and schools devoted to teaching and research in international relations were founded. Courses were organized and general textbooks on the subject began to appear. Private organizations were formed, and large grants of philanthropic funds were channelled to the support of scholarly journals, to the advancement of citizenship in world affairs through special training institutes, conferences, and seminars, and to the stimulation of university research.

Initially, three subject areas commanded the most attention. All three had roots in the period of World War I. In the revolutionary upheavals at the end of the war, great portions of the government archives of imperial Russia and imperial Germany were opened and made public in a series of documentary publications. Very exciting scholarly work began to appear that pieced together the theretofore-unknown history of prewar alliances, secret diplomacy, and military planning. These materials were integrated to provide explanations of the origins of World War I. The two decades between the two great wars were the heyday of diplomatic history, and the most famous of the students of international affairs were historians. With great ingenuity and industry, they presented the world with superb examples of the art and science of diplomatic history.

The second subject that captured attention was bound up with the hope and expectation of a new world order in the making through the League of Nations. Some of the schools of international relations that were founded in the

1920s had the explicit purpose of preparing civil servants for what was expected to be the dawning age of international government. Thus the genesis and organization of the League, the history of earlier plans for international federations, and the analysis of the problems and procedures of international organization and international law were investigated with enthusiasm.

The third study of consequence during the early part of the interwar period was an offshoot of the peace movement and was concerned with scholarly investigations of international warfare: its cause, its costs, and its sociological and psychological aspects. In addition to the data and the interpretations dredged up in the study of war, the interest in the question "why war?" brought a host of new social scientists—economists, sociologists, and psychologists—into active participation in international studies for the first time. They were pioneers in what later came to be known as the "behavioral approach" to international relations.

Realpolitik in the 1930s

The breakdown of the League, the rise of the aggressive dictatorships, and the coming of World War II in the 1930s caused a reaction against the international government and peace-inspired themes in the study of international relations. Idealism and moralism were criticized, and "realism" became the new thought in the field. The image was built at that time that the first stage of academic development of international studies was the handiwork of starry-eyed idealists and peace visionaries who ignored the hard facts of international politics. This characterization is untrue, the fact being that the scholarship on world affairs of the '20s and the '30s was extensive and sound in the organization of data and in the development of some fundamental concepts.

In the European tradition since early modern times, the knowledge of international relations had been loosely ordered in two branches of learning. The first is diplomatic history, which has been considered to reflect the variety of political experience, the particularity of events, and the contingencies in the actual practices of diplomacy and war. The second is international law, which has been viewed as registering the "residue of history"—the fundamental principles of conduct, the uniformities in international phenomena, and the permanent aspects of practice. The effect of the new field of international relations was to broaden the traditional organization almost beyond recognition.

New areas of inquiry during the interwar period

Some of the topics that today are considered novel and of recent origin were being explored vigorously in the two interwar decades; by the time of World War II, they already had acquired large bibliographies. It is instructive to recall a few of those topics in order to correct the stereotype that moralist teachings were then entirely dominant: the relationship of problems of racial and ethnic minorities to international affairs, the effects of the population explosion on foreign policies, the linkage between raw materials and other of the "life-support systems" of the planet with the actions of nations, the effects of imperialism and colonialism, the strategic aspects of international relations including the effects of geographical location and space on military power and the influence on governments of what has come to be called the "military-industrial complex," the economic inequalities of nations, and the role of public opinion, national differences, and cultural orientations in world affairs. If these studies tended to be short on theory and long on description, nevertheless the topics investigated remain relevant.

Certain individual scholarly contributions of the 1930s deserve particular notice because they were forerunners of what was to be developed after World War II. Harold D. Lasswell was making explorations of the relationships between world politics and the psychological realm of symbols, perceptions, and images. Abram Kardiner and his associates were laying the groundwork for a psychoanthropological approach to the analysis of national behaviour and culture, which later became a popular but short-lived theory of international relations. Frederick L. Schuman was producing foreign policy analyses that synthesized analytic comment with accounts of current international events. Schuman thus set the style that is still followed by government interpreters of foreign policy developments and by the news analysts of world affairs.

Quincy Wright was leading one of the first team research projects in the field and was investigating numerous aspects of international behaviour in a very broad approach to the study of war. Carl J. Friedrich, Frederick L. Schuman, Harold Sprout, Nicholas Spykman, E.H. Carr, Brooks Emeny, and others were developing the main lines of analysis of what became the power-politics explanation of international relations.

Some 30 years later, one begins to appreciate that the definition of the study of international relations and the widening of its scope were the fundamental contributions of the scholars of the interwar period. Many of the innovators of the 1930s found their services enlisted by governments during World War II for work in intelligence, propaganda, and political analysis. In this respect, the war stimulated systematic social-science investigations of international phenomena. On the other hand, World War II became a divide for academic international relations. The war made a drastic change in the agenda of world politics. The postwar intellectual climate shifted away from many of the earlier interests, emphases, and problems. There was a readiness in the early postwar years for an analysis that would cut through the details of studies of myriads of international topics and that would provide a focussed view of the fundamental nature of international politics. An intellectual hunger for theory existed.

The postwar ascendancy of political realism. Hans Morgenthau's *Politics Among Nations,* first published in 1948, met this need for theory. Writing in 1959, Stanley Hoffmann expressed what was, in all likelihood, the opinion of most students of international relations:

Hans Morgenthau and the "realist" theory of power politics

> The theory which has occupied the center of the scene in this country [the United States] during the last ten years is Professor Morgenthau's "realist" theory of power politics.

At the time of this writing, the influence of the Morgenthau text continues to be strong in most countries outside the Communist world. The realist theory still requires the attention of new students of international relations. A reader is best advised to explore the theory at its source. *Politics Among Nations* remains an impressive study; it is clearly conceived and well argued.

At the heart of the realist theory is the concept of interests. Politics is defined as the struggle for power, whether in domestic or international settings. The struggle for power is part of human nature and takes form in society according either to the competition or to the alignment of interests. Collaboration occurs when parties find their interests are coinciding. Rivalry, competition, and conflict result from the clash of interests. Accommodations are possible through the application of political skill.

In an international system composed of sovereign nation-states, the survival of both the states and the system depends on the intelligent pursuit of national interests and on the realistic calculations of national power. As long as the state system persists, the only truly constructive way to participate in international politics is through skilled diplomacy. Religious and ideological crusades threaten the ruin of both the individual states and the system, and disasters follow from attempts to reform nations toward the ideal of universal trust and cooperation.

Thus the realist theory of power politics was brought forward in the late 1940s to stand guard against idealists: those who would think and act in the visionary ways of moralism and legalism in world affairs. No impressive new formulation of political idealism appeared to carry the challenge to the realist position, and the "great debate" of realism versus idealism gradually faded from the scene.

Many scholars of international relations neither opposed nor accepted the power-politics theory. Some simply were engrossed in other aspects of international-relations teaching and research. Large sums of money were made available in the 1950s for the development of foreign-area studies, and general theoretical concerns played little part in the growth of area specialization. Other scholars agreed with Morgenthau's statement that theory and research should have a "concern with human nature as it actually is, and with the historic processes as they actually take

place," but they did not believe that the realist conceptualization provided a sufficient explanation of observed international behaviour.

The behavioral decade: mid-1950s to mid-1960s. An important new influence was the arrival in the field of a number of fresh ideas, conceptualizations, models, and paradigms that were loosely identified in ensemble as behavioral theory. The new movement distracted attention from the realist–idealist question. The unanticipated appearance in the mid-1950s of a large number of possible alternative ways to organize international data and to orient inquiries in international relations soon appeared to threaten the very foundations of scholarly communication. Simply to list a number of the conceptual innovations suggests the reason for the anxiety that the discipline might lapse into complete incoherence. In addition to power theory, there appeared a welter of theories, each with its distinctive label: decision making, system, conflict, deterrence, capabilities, field, communication, integration, development, environmental, cognitive, and, finally, game theory. Much of the intellectual effort of the "behavioral decade" went into the task of attempting to understand, compare, interpret, and integrate all these ideas. The scholarly goal of the period became to build an integrated framework of theories—to carry out conceptual mapping.

The behavioral aim to integrate the mass of ideas

To describe the efforts at theoretical integration that were made and the problems that were encountered would require book-length treatment. Suffice it to say that the comparing and integrating of the elements of theory turned out to be difficult. The more the matter was investigated, the more the specialists in theory questioned the necessity of arriving at one comprehensive structure of theory. The international realities that theory is supposed to reference and explain are varied and diverse, so why should one not expect that a number of separate theories would be needed to account for different parts and aspects of international relations?

Increasingly, explanations that trace the forces of international relations to any single source have been seen to be unsatisfactory. The struggle for power among nations, for example, can be accepted as a fact in past and current international politics, but to theorize that all other factors are subordinate to or dependent upon this one is to exclude too much that is important and interesting in international phenomena. Similarly, the formulation that asserts that the character of nations and, hence, the character of their participation in international relations are determined by child-rearing practices is simple, appealing; at the same time, however, it is unacceptable because it theorizes a single cause where multiple causation prevails. The Communist theory that international relations are the historical expression of the class struggle also falls into the single-source classification of theory.

The attitude nurtured in the behavioral decade emphasized the necessity of recognizing social multidimensionality and, therefore, multiple causation and the multiple forms of explanation and theory. Under this perspective, one might well conclude that the concept of the struggle for power among nations and the idea that international relations is really a manifestation of a global struggle between social classes both relate to human conflict and that a theory of conflict should encompass these as well as other conflict interpretations. Indeed, the anticipated integrating effect on theory and research in international relations was a main motive in advancing general conflict theory in the '60s, but it is important to add that conflict theory came to coexist with integration theory and game theory, both of which approach some conflict phenomena from different conceptual angles.

Development of the theory of conflict

By the end of the behavioral decade, the multiple-theory perspective had come into the ascendancy in North America and western Europe. International-relations scholars in the Soviet Union who were following closely the Western literature in the field reported with some satisfaction that Western theory had become eclectic and was in disarray in contrast to their own, which they declared to be based on the unified theory of the science of Marxism. At the same time, they gave indications that some of the Western trends in international-relations thinking were being introduced in socialist countries and that a muted contest between traditional and behavioral approaches was underway.

By the 1970s only the realist theory of power politics survived as a relatively simple and comprehensive explanation of international politics in a conceptual environment that otherwise had become pluralistic and complex to the extent that "theory" could no longer be outlined quickly or conveniently in a classroom or before an audience. A situation had begun to develop that put the question of a single theory of international relations in about the same class with that of a single theory of biology.

Contemporary perspectives in international relations. *Foreign-policy and international-systems perspectives.* One important clarification developed from the effort of the behavioral decade to bring the various theories together in a unified structure. It was the consensus that the academic organization of the discipline has two principal parts. It is convenient to call each part a perspective on the subject. The parts are called the foreign-policy and the international-system-analysis perspectives.

The foreign-policy perspective covers many theory and research interests. Fully generalized, it embraces all the inquiries that look into the domestic sources of external or international phenomena. Thus, a study of any set of traits, structures, or processes arising within a national society or polity that can be demonstrated to determine or influence importantly how that society or polity participates in international relations belongs to the foreign-policy perspective. The decision-making approach to international politics meets the requirement, for example. The analysis of the information that decision makers use, of their perceptions and motives, of the influences on them exerted by public opinion, and of the organizational settings in which they operate is a manifestation of the foreign-policy perspective. Studies that seek to relate the facts of wealth and power of a nation to its international status and role provide other illustrations.

Comparative foreign-policy analysis is an area of theory and research effort that first appeared in the mid-1960s. Its objectives are, first, to examine the data of domestic sources of external conduct country by country using standard criteria of data selection and analysis, and, then, to compare across countries for generalized findings on foreign-policy performances. When the details of the domestic sources and the external performances have been compared, theories about the domestic–external linkages and about the groupings of countries according to the types of linkage are expected to develop. The comparative foreign-policy approach, so described, develops theory through inductive research procedures.

The second perspective is that of international system analysis. Whereas foreign-policy analysis concentrates on the actors, international system analysis is preoccupied by the interaction. The term "interaction" suggests challenge and response, give and take, or move and countermove. Diplomatic histories feature the narratives of action and response in international situations and interpret the meanings in the exchanges. The theory of the balance of power is an example of an international-system conception. Explanations and descriptions of bargaining in international negotiations fit the perspective and so also do studies of arms races and other escalating processes. A model of the international trading system would be an example of a structural approach to international system studies, while an examination of how and why a coalition of states disintegrates would represent a process approach to international system analysis.

The theorist of international systems may gain a general outlook on the phenomena he studies through a system perspective, but he must bring in a certain amount of empirical detail when he identifies the components, relationships, and environments of the system that is the object of his theoretical inquiry. System theory does not turn out to be any single formulation; it is more in the nature of a conceptual anchoring point for a variety of specific formulations. Thus one system theorist may define the components of his system as geographical nation-states related according to rules and political structures, whereas another theorist may define the components of

his system as nonterritorial, nonstate transactional units only partly related by the influences wielded by national governments.

The general system perspective. Although the theoretical development of the system idea may lead to very diverse outcomes, another more general concept—that of open, adaptive systems—may provide the most promising approach to a comprehensive understanding of the dynamics of relations among nations. Without imposing any single school of thought or any single interpretation of world affairs, it has a loosely unifying effect on the outlook of students of the field. The so-called general system perspective on international relations may be compared to the map of a little-explored continent. Outlines, broad features, and a continental delineation are not in question, but everything else remains in doubt, is subject to controversy, and awaits exploration. One commentator has remarked that general system theory is not really a theory but instead is "a program or a direction in the contemporary philosophy of science." (From Anatol Rapoport, "Systems Analysis," in *International Encyclopedia of the Social Sciences,* vol. 15, p. 452, 1968.)

General concept of open, adaptive systems

As noted above, the quest for theoretical unification during the behavioral decade resulted in the widespread acceptance of two perspectives—a foreign-policy approach and an international-system approach. The general concept of open, adaptive systems provides a conceptual bridge connecting the two perspectives and creates a loose bond joining many of the diverse theoretical formulations prevailing in the field. An examination of the line of general thinking that builds the bridge and provides the bond is, therefore, well worth attention.

One begins to think about almost any open, adaptive system that involves human beings as a living system. If the system is living, its pervading characteristic is activity. Acting units, however they are recognized and defined, are doing things, participating in events, carrying forward processes, and creating effects. The effects created by activity include progressive influences on the actors. That is to say, the acting unit is immune neither to the effects of its own participation nor to the participatory influences generated by other acting units. It is this situation that establishes the condition of the openness of a system. Streams of influencing activity contain two kinds of processing: the first kind is regular, which is to say, governed by rules and repetitive in form, and the second kind is unexpected, irregular, and variant.

It is the second type that stimulates change and that initiates the special processing called adaptation. A living system is open and is able to adapt, however, only if it in some way has gained access to information on the state of affairs that joins it with its environment and, further, in some way has achieved means to direct and to change its stream of activity. Thus in addition to the fundamental concept terms of openness and adaptation, the general system perspective incorporates the ideas of communication and corrective action or, more generally, of communication and control.

Another view adds the observation that there is a systematic deception afflicting human understanding when the unit of action—the actor or the initiator—is seen as an entity. For social phenomena, the system perspective advises us that we make a grievous error when we identify individual persons, groups, organizations, nations, and so on as separate, uniquely named things. Recognized correctly, all these are nothing other than organized and interlocking activity flows, proceeding across historical time in somewhat regularized fashion. John Doe is to be described, literally, as an organized packet of active processes in exchange with an environment and utilizing communication and control to survive and adapt. The same description fits Japan: an incredibly complex network of related action much more than a place, a people, or a name.

Not only does the general system perspective urge that the "actor" be considered as a configuration of activity but it also prompts a recognition that, most often, the configuration itself has a hierarchical organization. In fact, the use of the term "system" is ordinarily an effort to convey the meaning that smaller organized activity flows serve larger activity flows and that the functional linking of subordinate parts to the operating whole is the process that defines what an actor really is. Thus the conventional explanation is that any recognizable living system is made up of related "components" and that each component, when examined at its own level, is found to be a functioning system in its own right. It also may be called a "subsystem." Enough of the system conceptualization has been suggested here to show next how these fundamentals have been translated for the purposes of international-relations theory.

In world history, the Earth has been populated by hundreds of separate and relatively isolated social systems. Each system, as a somewhat ordered stream of interrelated activities, has had exchanges with its particular environment in its own time and in its own way and has employed communication and control capabilities either to succeed at adaptation or to fail at it and, therefore, perish. What today is called international relations is that sector of exchange of a social system with its particular environment that has to do with interlinked action flows to and from other separate social systems. The knowledge problem of international relations is to understand, describe, and explain such flows to and from social systems from their source to their termination. Social systems thus consist of complexes of "internal" interacting subsystems or components, only some of which connect their processes with the action flows between that social system and another.

Internal processes and interactive processes of nations

The next logical thought then is that there can be only two sources of international conduct. One originates in the activity complexes within each participating social system and the other arises from the effects of the interlinked action flows to and from the participating social systems that together make up the membership of an international system. Hence, two basic perspectives of theory and research on international relations are distinguished by the primary attention given either to the origins of conduct arising from internal processes or to origins of conduct arising from the effects of the processes of exchange between social systems—or, to put it more succinctly, from the effects of interaction.

Obviously, the two perspectives are related because each casts its beam on a sector of a whole phenomenon of action and interaction. If he expects to be understood, the theorist must specify not only the basic perspective he wishes to emphasize but also particular identifications of the "units" of action, the kinds of action flows, and the linkages of processes and effects in this particular conception of the system that concerns him. In a multidimensional social world, the theorist can exercise his choice of a focus of inquiry in many different ways. Multiple theories are the outcome that may be expected from the introduction of the concept of general systems into the study of international relations.

The rise of quantitative research and computers. The emphasis on theory building and theory integrating after World War II was perhaps a reaction against the emphasis on methodology inherent in the historical and institutional studies of the 1930s. In turn, the renewed interest in empirical research after the behavioral decade probably was a reaction against the excessively intensive theorizing. Only part of the renewal of interest in data gathering and data analysis can be accounted for in this way, however. The marked increase in quantitative data studies after the mid-1960s has resulted from the direct influence of the computer. Computerization came late to the field of international relations; how to use the machines to good advantage was not apparent at first because most observations on the conditions of international relations are recorded in narrative or literary form.

Uses of computers in international relations

The first big discovery about computers in the study of international relations was that they could be made to serve the function of marvelously efficient librarians. By placing data collections under computer management, researchers could improve their investigations by incorporating vastly larger collections of facts in their work. Facts that had a narrative, nonquantitative form could be included in storage and retrieval systems almost as readily as numerical data. The second revelation was that by systematic coding

and counting, many kinds of nonquantitative data could be transformed into quantitative indicator information and then be condensed and evaluated by mathematical and statistical procedures. Vistas on new research possibilities were opened and exploratory studies were made in a number of directions. Through quantitative, computer-assisted research detailed, painstaking examination of historical records has again become important. The new research also gives promise of bringing the interests of academic theory and research into closer accord with the interests of government analysts and the practitioners of diplomacy.

RELATIONS BETWEEN SCHOLARSHIP AND ACTION IN FOREIGN AFFAIRS

The differences between the interests of scholars and practitioners of international affairs have appeared to be more prominent than the similarities. A steady concern of scholars has been to avoid both the fact and the reputation of serving as apologists for official foreign policies. One of the first teachings of the idealism of the discipline's founding period at the time of World War I was to maintain an attitude of future-oriented reform of the international order. Existing systems had lesser importance according to the early objectives of the field. The principle of scientific detachment in social-science research also has contributed to the scholarly effort to evaluate international events and developments from a global standpoint rather than from the perspective of any one country's foreign-policy position. On the other side, practitioners have been inclined more to indifference than to hostility in their attitudes toward academics in international relations. They frequently have professed that for their day-to-day work they have found little of value in the theory and research contributions of the field. One looks in vain, therefore, for many signs of direct influence in either direction. An indirect and subtle exchange has occurred, nevertheless, and it has had importance in the direction of foreign affairs.

Mutual influences of government and academia

New international programs or new foreign-policy directions undertaken by governments often have attracted so much interest in the universities that research programs have been initiated and special subfields of international studies have resulted. Examples of such subfields include: (1) "national development" stimulated by the foreign-assistance programs to aid underdeveloped countries; (2) the "area studies" that emerged from World War II because of the problems that Western governments had in finding knowledgeable personnel who understood the languages, histories, cultures, geography, and politics of Asian, African, and Latin-American regions; and (3) "national security," resulting from the heavy influence of military factors on foreign policy and especially from the nuclear-control problem in the post-World War II period.

The indirect influence flowing in the other direction from university studies into governmental thinking may be traced in a number of noteworthy examples during the second half of the 20th century: first, the realist formulation of the power-politics theory has filtered into the foreign-policy thinking of the U.S. government to such an extent that most foreign-policy decisions have been defended by arguments of national interest and power calculation and opposing views have been discounted for reflecting insufficient "hard-nosed" realism. Another interesting example was the preoccupation of Pres. John F. Kennedy's administration with the details of the process of foreign-policy decision making in crisis periods. This orientation can be traced to the popularity of decision-making theory advanced in the academic field in the preceding decade. One other example was the former Soviet Union's revision of foreign-affairs doctrine to accommodate the deterrence theory of nuclear defense. Concepts drawn from the literature of American "defense intellectuals" were readily evident in Soviet works on military doctrine.

Although the past relations of scholars and practitioners of international relations have been relatively mild and indirect, the future may change this customary relationship. The implications of system theory can be expected to percolate slowly but deeply into the understanding of both officials and the public. As this occurs, there will be greater appreciation of how truly complex and far-reaching the flows of international events are. The problems of survival and adaptation will be seen in a new light, and the realization of how interdependent modern nation-states are—how much they are bound to a common fate—should lead to more careful, better researched, and more deliberate formulations of foreign policy. The search for more accurate appraisals of the state of the world in its international aspect, which is the vocation of theory and research in academic international relations, should bring officials and researchers together.

Demand for data

On the other side of the relationship, the surge of growth in quantitative, computer-assisted studies in the universities is creating an increasing appetite for more frequent and more exact reports on the state of the world and in increasingly more of its aspects. The academic community in whatever country one might name lacks the resources and trained manpower to satisfy this growing appetite for data. The demand could well lead to increased responsiveness of the data-gathering services of national governments and to large reductions both in secrecy and in governmental interference with free public communication among nations. The current practice of withholding large amounts of information about developments in the flows of international events is an atavism left over from the era of aristocratic diplomacy. Today it has become a threat to the survival of humankind.

If the data on the conditions and relationships of the world's social systems (now made more manageable and more available for immediate use through computer systems) are brought into full and untrammeled circulation, the academic field of international relations and the governmental analysis and planning agencies will have much more in common than they have had. The end result should be the development of entirely new, more competent, and more realistic approaches to the formulation and execution of foreign policies. (C.A.McC./Ed.)

Comparative law

HISTORICAL DEVELOPMENT OF COMPARATIVE LAW

The expression comparative law is a modern one, first used in the 19th century when it became clear that the comparison of legal institutions deserved a systematic approach, in order to increase understanding of foreign cultures and to further legal progress. From early times, however, certain scholars and researchers have made use of the comparative technique, conscious of the advantages to be gained.

Ancient roots of law. In the 6th century BC according to legend, the Greek lawgiver Solon, faced with the task of compiling the laws of Athens, gathered together the laws of various city-states. Similarly, in the 5th century BC, a Roman commission was reported to have consulted the statutes of the Greek communities in Sicily before giving Rome the famous *Laws of the Twelve Tables.* Aristotle, in the 4th century, is said to have collated the constitutions of no fewer than 158 city-states in his effort to devise a model constitution. Thus, from ancient times it would seem that those wishing to set up a just system have sought inspiration and example from abroad. The true expansion of comparative law, however, was hindered by a number of obstacles—such as the parochialism of social groups, contempt for foreigners, or "barbarians," and belief in the sacredness or everlasting inviolability of inherited legal rules.

Although certain practices and institutions that crept into Roman law undoubtedly originated in the imperial provinces, Roman legal science took no cognizance of comparative law. Nor can the medieval universities in Europe be said to have displayed great concern for comparative law. Over the centuries, their interest was limited to Roman law, supplemented in certain areas or modified to some extent by canon law. While members of the first school of thought (called glossators) confined themselves to the task of elucidating the meaning of the Roman codes of law, their successors (the postglossators) undertook the systematic arrangement and adaptation of

The ancient search for "higher" or ideal law

that law to prevailing social conditions. At no time was there an effort to compare laws. The customary laws that one found here and there could hardly hold any interest for scholars labouring to give society a model of ideal justice and to discover or elucidate a higher law above man's making. Indeed, in their opinion, local laws were no more than rubbish and evidently doomed to decay. To compare these local practices would have been a waste of time; to compare them with Roman laws would have been almost indecent.

Role of judges. Such contempt was not characteristic of the attitude of the judges and lawyers whose duty it was to administer justice, mainly by applying the customary law. Their material contained areas of uncertainty and required adaptation to social needs. In the work of ascertaining the content of a custom, and in the task of filling the gaps of customs, judge or lawyer had to consider which customs to allow to prevail. In so doing, he had to decide whether one custom was more just than another and how far he should go in introducing concepts of ideal justice (based on Roman law) that were being promoted by the universities. Two processes were thus at work: the elimination of conflicting local customs and the acceptance and rejection of elements of Roman law. With regard to the first process, the comparative aspects of the work took place behind the scenes, and consequently the results of melding the different local or municipal laws are known, but the reasoning leading to the result is not. With regard to the second process, by contrast, certain publications place the act of comparison in full view. This was particularly noticeable in England, where some writers—such as Sir John Fortescue in the 15th century and Saint-Germain in the 16th—took upon themselves the comparison of common law and Roman law, and in 1623 Sir Francis Bacon suggested to James I that a work be drafted comparing English and Scots law, as a preliminary step toward the unification of the two systems.

19th-century beginnings. Despite the occasional use of the comparative technique, nevertheless, comparative law itself was not recognized as a separate branch or as a fundamental technique of legal science until the 19th century. In particular, it played no part in legal education. It was quite unthinkable that the pursuit of justice should be taught by reference to a host of customary rules that were incomplete, sometimes archaic, and generally regarded as barbaric. A foundation of ethical and political principles rather than sociological considerations, an appeal to reason rather than a study of human behaviour or judicial precedent—these were deemed the true criteria of progress.

End of dualism between ideal law and everyday practice

With the coming of the 19th century, codification of the law put an end to the dualism existing in many countries between an ideal system, as taught in the universities, and the laws that were applied in everyday practice. Codification of those everyday laws gave them the status of a national law, thoroughly purged of anachronisms and arranged in a systematic manner. That codified law became the cornerstone of legal education. This promotion of local customs, regarded henceforth as being fully consonant with natural justice, may be considered as the underlying cause of the appearance and rise of comparative law.

In short, the attitude toward comparative law tends to change when a country makes its national law the object of legal study and law students begin contrasting it with foreign counterparts. In Europe this dawning change was evident early in the 19th century. Legal periodicals were founded in Germany in 1829 and in France in 1834 to further a systematic study of foreign law. In France, the civil and mercantile laws of modern states were translated with "concordances" referring to the corresponding provisions of the French codes; and in England in 1850–52, Leone Levi published a work entitled *Commercial Law, Its Principles and Administration; The Mercantile Law of Great Britain Compared with Roman Law and the Codes or Laws of 59 Other Countries.*

A chair of comparative legislation was set up in 1831 in the Collège de France; and this was followed, in 1846, by a chair of comparative criminal law in the University of Paris. In 1869 the Société de Législation Comparée was founded in France, followed in 1873 by the Institut de Droit International and the International Law Association. In England, the Society of Comparative Legislation was founded in 1895, and the Quain Professorship of Comparative Law was created at London University in 1894. Similarly, chairs in comparative law were founded and projects in foreign law undertaken all over the continent of Europe, but with particular vigour in France.

First International Congress of Comparative Law

International efforts. The 19th century drew to a close with an important event—the meeting of the First International Congress of Comparative Law in Paris in 1900. Experts from every part of Europe delivered papers and discussed the nature, aims, and general interest of comparative law. Particular emphasis was laid on its role in the preparation of a "common law for the civilized world," the contents of which would be laid down by international legislation. The stress, however, was on comparative legislation and codification because (with the exception of one English jurist) the congress had attracted only jurists from continental European nations, all of which had coded law, in contrast to English customary, or common, law. Consequently, the idea of an enacted world law was the natural outcome of its proceedings.

The upheavals resulting from World War I (1914–18) prompted a change in direction. From then on, European interest began to extend beyond the continental systems themselves, first, to those of the common-law countries (chiefly England and the United States), then still further afield to the Socialist systems, and finally, after 1945, to the laws of the newly independent states of Asia and Africa. The new territory for legal study that has thus been opened up has resulted in references to comparative law, rather than to comparative legislation.

METHODOLOGICAL CONSIDERATIONS IN CONTEMPORARY COMPARATIVE LAW

The world contains a vast number of national legal systems. The United Nations brings together representatives of some 127 states, but these states are far outnumbered by legal networks, since not all states—notably federal ones—have accomplished unification within their own frontiers. It is thus an enormous task to try to compare the laws of all the different jurisdictions. This problem, however, should not be overly magnified. Differences between the diverse systems are not always of the same order; some are sharp; others are so closely similar that a specialist in one branch of a legal "family" often may easily extend his studies to another branch of that family. For this reason, one can distinguish two types of research in comparative law. The exponent of "microcomparison" analyzes the laws belonging to the same legal family. By observing their differences, he will decide whether they are justified and whether an innovation made in one country would have value if introduced elsewhere. The researcher pledged to "macrocomparison," on the other hand, investigates those systems differing most widely from each other in order to gain insight into institutions and thought processes that are foreign to him. For the "pure jurist," concerned mainly with legal technicalities, microcomparison holds the greater attraction; whereas macrocomparison is the realm of the political scientist or legal philosopher, who sees law as a social science and is interested in its role in government and the organization of the community.

Two types of research in comparative law

Microcomparison. Microcomparison demands no particular preparation. The specialist in one national system is usually qualified to study those of various other countries of the same general family. His chief need is access to bibliographical material. In the United States, each state has its own statutes and, to some purposes, its own common law. Thus, the American lawyer must be a microcomparatist as he takes the 50 state systems and the federal law into daily account in his practice of the law. The same is true, to a large extent, of the Australian, or Indian, or Kenyan lawyer, who must take into account not only his own national system but also the laws of England and of other common-law jurisdictions in the Commonwealth. Whatever can be said of the common-law systems holds largely true for the Roman-law and Socialist families. French comparative law students encounter little difficulty in contrasting the laws of certain countries, so long as they

confine their study to French, German, Italian, and Dutch law, which are related in tradition and structure and serve a similar type of society.

Macrocomparison. The situation differs greatly in consideration of macrocomparison. Here no comparison is possible without previously identifying and thoroughly mastering the fundamentals of the law systems as they differ from place to place. The jurist must, as it were, forget his training and begin to reason according to new criteria. If he is French, English, or American, he must recognize that in some folk societies of the Far East, the upright citizen never crosses the threshold of a courtroom and acknowledges no subjective rights; instead, the citizen's behaviour is governed by rites handed down from his ancestors, ensuring him the approval of the community. Likewise, if the Western jurist is to understand the law of the Muslim or Hindu, he must realize that the law is contained in rules of conduct laid down by a religion for its followers, and for its followers only. These rules, creating obligations and not rights, rank above all worldly matters and, in particular, are not to be confused with the regulations that a national government may, at a given time, enact and ratify. Further, in comparing his system of law with that of a Communist nation, the Westerner must remember that on no account does the citizen of a Marxist-Leninist state regard the rule of law as an ideal for society. Far from it, for his dream is to see law—which to him is synonymous with injustice and coercion—wither away in an affluent society founded on human solidarity and fellowship. A considerable shifting of legal gears is necessary before a French or German jurist can grasp the vital importance that the English or American lawyer traditionally attaches to the concept of due process and the rules of evidence; in continental eyes, procedural rules take second place to substantive law.

Mastering fundamentals of a foreign system

The specialist of macrocomparison also picks out the structural differences existing between certain systems. Accordingly, the Anglo-American lawyer must be aware of the importance of the distinction between public and private law—between law involving the state and law involving only individuals. The jurist in a Roman-law country must, conversely, appreciate the significance of the concepts of common law (unwritten customary law of various kinds) and equity (the use of injunctions and other equitable remedies), neither of which have counterparts in his own system. The lawyer from a centralized country must familiarize himself with the distinction between federal law and the laws of secondary jurisdictions (states, provinces, cantons, and so forth)—a distinction that is of fundamental importance in many countries. If he is from a nation like England or France that acknowledges the sovereignty of the national parliament, he must give due weight to the prominence of constitutional law in countries that permit courts to review the constitutional validity of legislative acts—especially in countries such as the United States and the Federal Republic of Germany. The jurist in a "bourgeois" country must appreciate the policy of collective ownership of means of production in Socialist states.

Parliamentary versus constitutional law

Classification of families of law. The terms microcomparison and macrocomparison, reflecting the language of economics, are in keeping with the idea that legal systems can be grouped into families, such as common-law, Roman, and Socialist. But it must be acknowledged that the number of identifiable families and the appropriate classification of a given system are questions always open to argument. The legal system of a given country, for instance, may exhibit some features that relate it to a particular family and others that may escape that classification. Such blurring of distinctions is particularly true of law in countries of Africa and the Middle East, where certain sectors of the law have been transformed by Western ideas (as in criminal and mercantile law and procedure) leaving other sectors (such as personal status, family law, and land law) faithful to traditional principles of the region. The phenomenon is not peculiar to those countries, however.

Wide differences also may be detected between legal systems that are commonly regarded as belonging to the same family. American law, for instance, without hesitation is ranked as a member of the common-law family; yet countless differences set it apart from English law, in large part because the United States has a federal and England a unitary system of government.

PURPOSES OF COMPARATIVE LAW

Historical and cultural comparisons. First of all, there has been a tendency to view comparative law from the standpoint of its value to the historical study of legal decision making—a consideration that was responsible for establishing the first chairs of comparative law in 19th-century Europe. Ideas regarding the place of law in society and the nature of the law itself—whether divine or secular, whether dealing with substantive or procedural rules—obviously become appreciably clearer when comparative law is joined to historical research. Indeed, to some extent historical background may aid in forecasting the future of certain national systems

A closely related consideration prompts many Western jurists, political scientists, and sociologists to acquaint themselves with non-Western methods of reasoning. Comparative studies reveal that the citizen of some countries of Asia and Africa looks upon the concept of a just social order with thoughts and feelings far removed from those of Western man. The notions of a rule of law and of rights of the individual—fundamental to Western civilization—are not wholly recognized by those societies that, faithful to the principle of conciliation and concerned primarily with harmony within the group, do not favour excessive Western-style individualism or the modern Western ideal of legal supremacy. Thus comparative law may enable statesmen, diplomats, and jurists to understand foreign points of view, and it may frequently help to create better international understanding.

Understanding another culture's reasoning

Commercial uses. Comparative law may be used for essentially practical ends. The businessman, for instance, needs to know what benefits he may expect, what risks he may run, and generally how he should act if he intends to invest capital or make contracts abroad. It was with this purpose in mind that the first French institute of comparative law was set up in Lyon in 1920; its mission was to instruct French legal advisers on foreign trade. It was this practical aspect that also encouraged the growth of comparative law in the United States, where the essential aim of the law school has been usually to turn out practitioners; and one need hardly mention the strong link in Germany between big industry and the various institutes of comparative law. Sometimes it is said that studies with such a focus should not be considered a part of comparative law, but practical considerations certainly have helped to finance and promote the development of comparative legal studies in general.

Aid to national law. The improvement of national legislation was the prime consideration during the 19th century in countries that were codifying or recodifying their legal systems. Numerous later additions to the Code Napoléon, drawn up in 1804, for instance, were of foreign origin. Many other nations, of course, followed France's lead and introduced into their own systems elements of the French Napoleonic codes and institutions of French public law. It is well worth noticing that a book on French administrative laws was published in German by Otto Mayer before Mayer felt himself able to write a textbook on German administrative law.

The foreign inspiration of a number of legal rules or institutions is a well-known phenomenon, sometimes so all-embracing that one speaks of "reception"—reception, for instance, of the English common law in the United States, Canada, Australia, India, and Nigeria; reception of French law in French-speaking Africa, Madagascar, Egypt, and Indochina; reception of Swiss law in Turkey; and reception of both German and French law in Japan, along with even some reception of American common law. The study of comparative law has found a special place in countries where such a reception has occurred.

Use in international law. In modern times the spirit of nationalism has often tended to frustrate the development of an international law that would overcome individual national differences. One task facing statesmen and jurists is to inject new life into this effort, adapting it to the

exigencies of the modern world. Those engaging in international trade, for instance, do not know with certainty which national law will regulate their agreements, since the answer depends to a large extent on a generally undecided factor—namely, which national court will be called upon to decide the questions of competence. Thus, the sole lasting remedy would seem to be the development of an international law capable of governing all legal questions outside the jurisdiction of a single state. Such a project can succeed only through the medium of comparative law.

(R.Da./Ed.)

BIBLIOGRAPHY. PRESERVED SMITH, *A History of Modern Culture*, 2 vol.(1930–34, reprinted 1962), covering the years 1543–1776, is a classic in the history of ideas and the best single work on the period leading up to the emergence of the individual social sciences. JAMES WESTFALL THOMPSON, *A History of Historical Writing*, 2 vol. (1942), is useful in this respect also. On the age immediately preceding the rise of the social sciences, the best study by far is LESTER G. CROCKER, *Nature and Culture: Ethical Thought in the French Enlightenment* (1963), and *An Age of Crisis: Man and World in Eighteenth Century French Thought* (1959), recommended to be read or consulted in that order. The best general work on the history of the social sciences and the history of social philosophy in the West is HARRY ELMER BARNES and HOWARD BECKER, *Social Thought from Lore to Science*, 2nd ed., 2 vol. (1952).

ERIC J. HOBSBAWM, *The Age of Revolution: 1789–1848* (1962), is an important and fascinating treatment of the social, cultural, and intellectual aspects of the age in which the individual social sciences emerged in western Europe. ROBERT A. NISBET, *The Sociological Tradition* (1966), although concerned primarily with sociology, deals with the specific ways in which the ideologies and themes of the democratic and industrial revolutions became translated into social theory. The same author's *Social Change and History* (1969) deals in detail with the incorporation of the theory of social evolution into the social sciences of the 19th century. For the rise and development of the individual social sciences in the 19th and 20th centuries, the following works are recommended. (*Anthropology*): ROBERT H. LOWIE, *The History of Ethnological Theory* (1937); and MARVIN HARRIS, *The Rise of Anthropological Theory: A History of Theories of Culture* (1968). (*Economics*): ERICH ROLL, *A History of Economic Thought*, 3rd ed. rev. (1954); and the extremely readable ROBERT HEILBRONER, *The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers*, 3rd ed. (1967). (*Political science*): GEORGE SABINE, *A History of Political Theory*, 3rd ed. (1959), best on the three centuries preceding the 20th; FRANCIS W. COKER, *Recent Political Thought* (1934), excellent for the early 20th century; and LEO STRAUSS and JOSEPH CROPSEY (eds.), *History of Political Philosophy* (1963). (*Sociology*): Barnes and Becker, referred to above, for detailed information on the history of sociology in the 19th and early 20th centuries; Nisbet, also referred to above, dealing with the relation between political ideologies and the currents of sociological thought in the late 19th century; and LEWIS A. COSER, *Masters of Sociological Thought* (1971), a very good general history of sociology in 19th- and 20th-century Europe and America. (*Social psychology*): FAY BERGER KARPF, *American Social Psychology: Its Origins, Development and European Background* (1932), the best account of social psychology in the 19th and early 20th centuries, and THOMAS C. WIEGELE, *Biology and the Social Sciences* (1982), concerning the effect of biological research on social science disciplines.

Anthropology: Histories of anthropological science include T.K. PENNIMAN, *A Hundred Years of Anthropology*, 3rd ed. rev. (1965), which covers all of anthropology; and P. MERCIER, *Histoire de l'anthropologie* (1966), which covers only cultural anthropology. The principal textbooks are M.J. HERSKOVITS, *Man and His Works* (1948); F.M. KEESING, *Cultural Anthropology: The Science of Custom* (1958); and J. POIRIER (ed.), *Ethnologie générale* (1968). In cultural anthropology—aside from two works by the "fathers" of the discipline, L.H. MORGAN, *Ancient Society* (1877); and E.B. TYLOR, *Anthropology: An Introduction to the Study of Man and Civilisation* (1881)—some of the classic general works are FRANZ BOAS, *The Mind of Primitive Man* (1911), and *Race, Language and Culture* (1940); BRONISLAW MALINOWSKI, *A Scientific Theory of Culture, and Other Essays* (1944); A.R. RADCLIFFE-BROWN, *Structure and Function in Primitive Society* (1952); A.L. KROEBER (ed.), *Anthropology Today* (1953); CLAUDE LEVI-STRAUSS, *Anthropologie structurale* (1958; Eng. trans. 1963); G. BALANDIER, *Anthropologie politique* (1967; Eng. trans. 1971); and M. MAUSS, *Oeuvres* (1968). Studies of individual peoples that have become classics include W.H.R. RIVERS, *The Todas* (1906); M. GRANET, *Fêtes et chansons anciennes de la Chine* (1919; Eng. trans., *Festivals and Songs of Ancient China*, 1932); BRONISLAW MALINOWSKI, *The Argonauts of the Western Pacific* (1922), and *Coral Gardens and Their Magic* (1935); A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922); MARGARET MEAD, *Coming of Age in Samoa* (1928); FRANZ BOAS, *The Religion of the Kwakiutl Indians* (1930); R.F. FORTUNE, *Sorcerers of Dobu: The Social Anthropology of the Dobu Islanders of the Western Pacific* (1932); R.W. FIRTH, *We, the Tikopia: A Sociological Study of Kinship in Primitive Polynesia* (1936); M.J. HERSKOVITS, *Dahomey: An Ancient West African Kingdom* (1938); E.E. EVANS-PRITCHARD, *The Nuer* (1940); and E.R. LEACH, *Political Systems of Highland Burma: A Study of Kachin Social Structure* (1954).

Sociology: Among older titles of major significance, generally considered classics, are W.G. SUMNER, *The Folkways* (1907, reissued 1940); C.H. COOLEY, *Human Nature and the Social Order* (1902, reissued 1967); G.H. MEAD, *Mind, Self, and Society*, ed. by C.W. MORRIS (1934); E. FARIS, *The Nature of Human Nature* (1937); R.E. PARK and E.W. BURGESS, *Introduction to the Science of Sociology* (1921, reissued 1969); and W.F. OGBURN, *Social Change with Respect to Culture and Original Nature*, new ed. (1950; suppl. ch., 1964). A comprehensive summary of early sociological theory is available in H.E. BARNES (ed.), *An Introduction to the History of Sociology* (1948). The following titles provide excellent coverage of the main directions and subfields of contemporary sociology: R.E.L. FARIS (ed.), *Handbook of Modern Sociology* (1964); R.K. MERTON, L. BROOM, and L.S. COTTRELL (eds.), *Sociology Today*, 2 vol. (1959, reissued 1965); G.D. GURVITCH and W.E. MOORE (eds.), *Twentieth Century Sociology* (1945); J.G. MARCH (ed.), *Handbook of Organizations* (1965); and N.J. SMELSER and J.A. DAVIS (eds.), *Sociology* (1969). Two influential general texts are L. BROOM and P. SELZNICK, *Sociology*, 4th ed. (1968); and G.A. LUNDBERG *et al.*, *Sociology*, 4th ed. (1968). Leading books treating communities and societies as wholes are TALCOTT PARSONS, *Societies: Evolutionary and Comparative Perspectives* (1966); I.T. SANDERS, *The Community*, 2nd ed. (1966); R.K. MERTON, *Social Theory and Social Structure*, rev. ed. (1957); G.C. HOMANS, *Social Behaviour: Its Elementary Forms* (1961); G.E. LENSKI, *Human Societies: A Macrolevel Introduction to Sociology* (1970); and A.L. STINCHCOMBE, *Constructing Social Theories* (1968). The technical and statistical methods used in sociology are presented in E.F. BORGATTA (ed.), *Sociological Methodology* (1968); and J.H. MUELLER, K.F. SCHUESSLER, and H.L. COSTNER, *Statistical Reasoning in Sociology*, 2nd ed. (1970). ERNEST GELLNER, *Soviet and Western Anthropology* (1980), illustrates differing approaches to anthropological study.

Social psychology: Excellent general textbooks include EDWIN P. HOLLANDER, *Principles and Methods of Social Psychology*, 2nd ed. (1971); and ROGER W. BROWN, *Social Psychology* (1965). A comprehensive account of research is G. LINDZEY and E. ARONSON (eds.), *Handbook of Social Psychology*, 2nd ed., 5 vol. (1968–69). A useful account of theories is MARVIN E. SHAW and PHILIP R. CONSTANZA, *Theories of Social Psychology* (1970). Social psychology approached through detailed analysis of social interaction is described in MICHAEL ARGYLE, *Social Interaction* (1969). Research on social psychology in industry is described in BERNARD M. BASS, *Organizational Psychology* (1965). Social behaviour in relation to personality is dealt with in EDGAR F. BORGATTA and WILLIAM W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968). A so-called symbolic interactionist approach is represented by GREGORY P. STONE and HARVEY A. FARBERMAN, *Social Psychology Through Symbolic Interaction* (1970); and by ERVING GOFFMAN, *Relations in Public* (1971).

Criminology: Important titles include H. MANNHEIM, *Comparative Criminology*, 2 vol. (1965); M.E. WOLFGANG and F. FERRACUTI, *Il comportamento violento* (1966; Eng. trans., *The Subculture of Violence*, 1967); H. MANNHEIM (ed.), *Pioneers in Criminology*, 2nd ed. (1972); L. RADZINOWICZ, *Ideology and Crime* (1966), a survey containing much historical material; J.T. SELLIN and M.E. WOLFGANG, *The Measurement of Delinquency* (1964), an attempt to find an operational definition of serious offenses and to construct a "crime index"; H. MANNHEIM, *Social Aspects of Crime in England Between the Wars* (1940); F.H. MCCLINTOCK and N.H. AVISON, *Crime in England and Wales* (1968), partly an updating of the preceding work; L.T. WILKINS, *Social Deviance* (1964), an attempt to bridge the gap between social research and social action from the viewpoint of the statistician; H. MANNHEIM and L.T. WILKINS, *Prediction Methods in Relation to Borstal Training* (1955); S. and E. GLUECK, *Predicting Delinquency and Crime* (1959), and *Ventures in Criminology* (1964); and A.K. BOTTOMLEY, *Criminology in Focus: Past Trends and Future Prospects* (1979).

Economics: The best introduction to methodological issues in economics is still L. ROBBINS, *The Nature and Significance of Economic Science*, 2nd ed. (1935). A more modern position on empirical testing is well conveyed by M. FRIEDMAN in *Essays in Positive Economics* (1953). The best way to learn about

economics is to browse through an introductory textbook, such as P.A. SAMUELSON, *Economics: An Introductory Analysis,* 8th ed. (1970); R.G. LIPSEY and P.O. STEINER, *Economics,* 2nd ed. (1969); or A.A. ALCHIAN and W.R. ALLEN, *University Economics* (1964). Good histories of economic thought include H.L. HEILBRONER, *The Worldly Philosophers,* rev. ed. (1953), a pleasure to read; O.H. TAYLOR, *A History of Economic Thought* (1960), arranged chronologically; and E. WHITTAKER, *A History of Economic Ideas* (1940), arranged topically. J.A. SCHUMPETER, *History of Economic Analysis* (1954); and M. BLAUG, *Economic Theory in Retrospect,* 2nd ed. (1968), are advanced references. Excellent articles on the great economists, as well as superb but sometimes quite difficult essays on the leading branches of modern economics, may be found in the *International Encyclopedia of the Social Sciences,* 16 vol. (1968). Each branch of economics has its own specialized texts. (*Microeconomics*): C.E. FERGUSON, *Microeconomic Theory,* 2nd ed. (1969); K.J. COHEN and R.M. CYERT, *Theory of the Firm: Resource Allocation in a Market Economy* (1965). (*Macroeconomics*): G. ACKLEY, *Macroeconomic Theory* (1961); T.F. DERNBURG and D.M. MCDOUGALL, *Macroeconomics,* 3rd ed. (1968). (*Development economics*): C.P. KINDLEBERGER, *Economic Development,* 2nd ed. (1965); B.H. HIGGINS, *Economic Development,* 2nd ed. (1968). (*Public finance*): R.A. MUSGRAVE, *The Theory of Public Finance* (1959). (*Monetary economics*): A.G. HART and P.B. KENEN, *Money, Debt and Economic Activity,* 3rd ed. (1961); L.V. CHANDLER, *The Economics of Money and Banking,* 5th ed. (1969). (*International economics*): C.P. KINDLEBERGER, *International Economics,* 4th ed. (1968). (*Labour economics*): A.M. CARTTER and F.R. MARSHALL, *Labour Economics: Wages, Employment, and Trade Unionism* (1966); L.C. HUNTER and D.J. ROBERTSON, *Economics of Wages and Labour* (1969). (*Industrial organization*): E.H. CHAMBERLAIN (ed.), *Monopoly and Competition and Their Regulation* (1954); J.S. BAIN, *Industrial Organization* (1959). (*Agricultural economics*): T.W. SCHULTZ, *The Economic Organization of Agriculture* (1953), and *Transforming Traditional Agriculture* (1964). (*Growth economics*): R.G.D. ALLEN, *Macro-Economic Theory: A Mathematical Treatment* (1967), a fairly difficult book; most texts on macroeconomics and economic development devote a chapter or two to growth theory. (*Mathematical economics*): A.C. CHIANG, *Fundamental Methods of Mathematical Economics* (1967). (*Econometrics*): A.A. WALTERS, *An Introduction to Econometrics* (1968); C. CHRIST, *Econometric Models and Methods* (1966). T.W. HUTCHISON, *The Politics and Philosophy of Economics: Marxians, Keynesians and Austrians* (1981), a history of economic thought; and DANIEL BELL and IRVING KRISTOL (eds.), *The Crisis in Economic Theory* (1981).

Political science: Although works of classical political philosophy are both venerable and extensive, few of them qualify as modern political science, because they are neither quantitative nor, in most respects, even empirical in tone and temper. ARISTOTLE's *Politics* and MACHIAVELLI's *The Prince* come closest to meeting empirical standards. AUGUSTE COMTE, *Cours de philosophie positive,* 6 vol. (1830–42; Eng. trans., *The Positive Philosophy of Auguste Comte,* 2 vol., 1853), and *Système de politique positive,* 4 vol. (1851–54; Eng. trans., *System of Positive Polity,* 4 vol., 1875–77), are seminal statements in the 19th century on a science of society. LUDWIG GUMPLOWICZ, *Grundriss der Sociologie* (1885; Eng. trans., *The Outlines of Sociology,* 1899; 2nd ed., 1963); and GUSTAV RATZENHOFER in *Wesen und Zweck der Politik,* 3 vol. (1893), argue the case for the primacy of groups in studies of the state. A useful summary statement of the sociologies of the 19th century is NICHOLAS S. TIMASHEFF, *Sociological Theory: Its Nature and Growth,* 3rd ed. (1967). A good general work on the efforts of German jurists in the 19th century to cope with the facts of federalism is RUPERT EMERSON, *State and Sovereignty in Modern Germany* (1928).

The most notable precursor of the behavioral approach in the 20th century was ARTHUR F. BENTLEY, *The Process of Government: A Study of Social Pressures* (1908, reprinted 1949). Others were GRAHAM WALLAS, *Human Nature in Politics,* 4th ed. (1962); and WALTER LIPPMANN, *Public Opinion* (1922; paperback ed., 1965). Besides works of the Chicago School mentioned in the article, the following may be noted: CHARLES E. MERRIAM, *Chicago: A More Intimate View of Urban Politics* (1929, reprinted 1970); LEONARD D. WHITE, *The Prestige Value of Public Employment in Chicago* (1929); and HAROLD D. LASSWELL and DANIEL LERNER (eds.), *The Policy Sciences: Recent Developments in Scope and Method* (1951), an effort to bring scientific method to the study of choices in public policy. Support for the establishment of a value-free science of politics was also provided by STUART A. RICE, *Quantitative Methods in Politics* (1928, reprinted 1969), who wrote the first general work on the application of statistical methods to the study of politics; GEORGE E.G. CATLIN, *The Science and Method of Politics* (1927); and WILLIAM BENNETT MUNRO, *Invisible Government* (1928). A useful summary survey of political science around the world after the end of World War II is *Contemporary Political Science,* published in 1950 by the UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION.

International relations: ADDA B. BOZEMAN, *Politics and Culture in International History* (1960), is the best guide to histories and concepts of past international systems. Other useful references that survey premodern international relations are COLEMAN PHILLIPSON, *The International Law and Custom of Ancient Greece and Rome,* 2 vol. (1911); RICHARD L. WALKER, *The Multi-State System of Ancient China* (1953); GEOFFREY F. HUDSON, *Europe and China: A Survey of Their Relations from the Earliest Times to 1800* (1931); FRANK M. RUSSELL, *Theories of International Relations* (1936); and SHMUEL N. EISENSTADT, *Political Systems of Empires* (1963). Accounts of the earlier development of the academic study of international relations may be found in SIR ALFRED ZIMMERN, *The Study of International Relations* (1931); and GRAYSON L. KIRK, *The Study of International Relations in American Colleges and Universities* (1947). Representative of the work of the 1930s that widened the scope of the field are HAROLD D. LASSWELL, *World Politics and Personal Insecurity* (1935; paperback ed., 1965); FREDERICK L. SCHUMAN, *International Politics: An Introduction to the Western State System and the World Community,* 6th ed. (1958); C.K. LEITH, *World Minerals and World Politics* (1931, reprinted 1970); W.S. THOMPSON, *Danger Spots in World Population* (1929); PAUL RADIN, *The Racial Myth* (1934); NICHOLAS SPYKMAN, *America's Strategy in World Politics* (1942, reprinted 1970); CARL J. FRIEDRICH, *Foreign Policy in the Making* (1938); BROOKS EMENY, *The Strategy of Raw Materials* (1935); ABRAM KARDINER, *The Psychological Frontiers of Society* (1945); QUINCY WRIGHT, *A Study of War,* 2nd ed., 2 vol. (1965); and E.H. CARR, *The Twenty Years' Crisis, 1919–1939,* 2nd ed. (1946, reprinted 1964). The theory of political realism is expressed by HANS J. MORGANTHAU in *Politics Among Nations* (1948). The idealist attempt to answer the challenge of political realism may be traced in THOMAS I. COOK and MALCOLM MOOS, *Power Through Purpose: The Realism of Idealism as a Basis for Foreign Policy* (1954); and JOHN H. HERZ, *Political Realism and Political Idealism* (1951). Four books on the psychological and cultural aspects of the behavioral study of international relations providing an excellent introduction are J. DAVIS SINGER, *Human Behavior and International Politics* (1965); OTTO KLINEBERG, *The Human Dimension in International Relations* (1964); JOSEPH H. DE RIVERA, *The Psychological Dimension of Foreign Policy* (1968); and HERBERT C. KELMAN, *International Behavior* (1965). The most famous book of its time on a psychocultural interpretation of national behaviour was RUTH BENEDICT, *The Chrysanthemum and the Sword* (1946, reprinted 1967). RICHARD C. SNYDER, H.W. BRUCK, and BURTON SAPIN, *Decision-Making as an Approach to the Study of International Politics* (1954), is still basic reading on the decision-making approach, but see also the data in GLENN D. PAIGE, *The Korean Decision* (1968). The most convenient compilation of varied examples of the theory and research of the behavioral decade is JAMES N. ROSENAU, *International Politics and Foreign Policy,* rev. ed. (1969). For conflict theory and international applications of game theory, see KENNETH E. BOULDING, *Conflict and Defense: A General Theory* (1962); THOMAS C. SCHELLING, *The Strategy of Conflict* (1960); and ANATOL RAPOPORT, *Fights, Games and Debates* (1960). EDWARD MCWHINNEY, *Conflict and Compromise: International Law and World Order in a Revolutionary Age* (1981), studies the effectiveness of law in resolving disputes among nations.

An introduction to general system theory in the social sciences is WALTER BUCKLEY, *Sociology and Modern Systems Theory* (1967). MORTON A. KAPLAN, *System and Process in International Politics* (1957); and CHARLES A. MCCLELLAND, *Theory and the International System* (1966), take different approaches to the applications of general systems ideas in the study of international relations.

Comparative law: A good historical survey is H.C. GUTTERIDGE, *Comparative Law,* 2nd ed. (1949). J.H. WIGMORE, *A Panorama of the World's Legal Systems,* 3 vol. (1928; 1-vol. ed., 1936), gives a general account of 16 legal systems. The ASSOCIATION OF AMERICAN LAW SCHOOLS, *A General Survey of Events, Ideas, Persons and Movements in Continental Legal History* (1912), is a pioneer book, in which the legal traditions of various European countries are discussed by specialists from each country. R. DAVID and J.E.C. BRIERLEY, *Les Grands systèmes de droit contemporains* (1964; Eng. trans., *Major Legal Systems in the World To-day,* 1968), a more up-to-date book originally devised for students, describes the problems and value of comparative law and provides information on civil, Socialist, common, and religious and traditional law, ending with a valuable bibliography. *The International Encyclopaedia of Comparative Law,* proposed 16 vol. (1971–), will, when completed, constitute a major source of information.

Social Structure and Change

Social structure and social change are general concepts used by social scientists, particularly in the fields of sociology and social and cultural anthropology. They are often conceived of as polarized twin concepts, social structure referring to permanence, social change to the opposite. The relationship between the two concepts is, however, more complicated. "Structure," for instance, does not necessarily indicate lack of change. Those features of a society, or any other social group, that are regarded as parts of its structure are always generated by dynamic processes. For example, the kinship structure of a given society (the typical composition of household units and the rules governing marriage and line of descent) is maintained by continuous changes in families, as marriages are concluded; children are born, grow up, and become adults; and people die. Second, although many social processes show a cyclical pattern—the formation, dissolution, and reformation of families being one example—social life never repeats itself completely. The kinship relations in one generation are never an exact replica of those in the previous one. The same processes that serve to maintain the social structure may also lead to social change and modification of the structure over a long period.

The concepts of social structure and social change pertain not only to basic characteristics of human social life but also to certain ideals and preferences. The structure, or order, of the society, generally regarded as harmonious and conducive to the general well-being, has also been seen as conflict-ridden and repressive. Similarly, social change has been conceived of both as progress and as decay, as emancipation on the one hand and as deviance from good tradition on the other. Such widely varying evaluations have influenced different theories concerning the nature of social structure and social change, and they continue to be reflected, to some extent, in present-day social thought.

This article is divided into the following sections:

Social structure 365
  Structural functionalism
  Theories of class and power
  Structuralism
Social change 366
  Historical background
  Patterns of social change
  Explanations of social change
  Mechanisms of social change
Conclusion 370
Bibliography 371

SOCIAL STRUCTURE

The term structure has been used with reference to human societies since the 19th century. Before that time, it had been already applied to other fields, particularly construction and biology. Its biological connotations are evident in the work of several social theorists of the 19th and early 20th centuries, such as Herbert Spencer in England. He and others conceived of society as an organism, the parts of which are interdependent and thereby form a structure that is similar to the anatomy of a living body.

Structure compared to anatomy

The metaphor of construction is clear in the work of Karl Marx, where he speaks of "the economic structure [*Struktur*] of society, the real basis on which is erected a legal and political superstructure [*Überbau*] and to which definite forms of social consciousness correspond." This phrase expresses the Marxian view that the basic structure of society is economic, or material, and determines, at least to a large extent, the rest of social life, which is defined as spiritual or ideological.

Although social scientists since Spencer and Marx have disagreed on the concept of social structure, their definitions have certain elements in common. In the most general way, social structure may be defined as those features of a social entity (a society or group within a society) that have a certain permanence over time, are interrelated, and determine or condition to a large extent both the functioning of the entity as a whole and the activities of its individual members.

Ideas implicit in structure

As may be inferred from this definition, several ideas are implicit in the notion of social structure. The concept expresses the idea that human beings form social relations that are not arbitrary and coincidental, but exhibit some regularity and persistence. The concept also refers to the observation that social life is not amorphous but is differentiated into groups, positions, and institutions that are interdependent, or functionally interrelated. These differentiated and interrelated characteristics of human groupings, although constituted by the social activities of individuals, are not a direct corollary of the wishes and intentions of these individuals; instead, individual choices are shaped and circumscribed by the social environment. The notion of social structure implies, in other words, that human beings are not completely free and autonomous in choosing their activities, but rather they are constrained by the social world they live in and the social relations they form with one another.

The social structure is sometimes simply defined as patterned social relations—those regular and repetitive aspects of the interactions between the members of a given social entity. Even on this descriptive level, the concept is highly abstract: it selects only certain elements from ongoing social activities. The larger the social entity considered, the more abstract the concept tends to be. What is considered as the social structure of a small group is generally much nearer to the daily activities of its individual members than that which is regarded as the social structure of a larger society. In the latter case the problem of selection is acute: what to include or not include as components of the social structure. The solution to the problem varies with the different theoretical views according to which characteristics of the society are regarded as particularly important.

Aspects of structure

Apart from these different theoretical views, some preliminary remarks on general aspects of the social structure of any society may be made. Most generally, social life is structured along the dimensions of time and space. Specific social activities take place at specific times, and time is divided into periods that are connected with the rhythms of social life—the routines of the day, the month, and the year. Specific social activities are also organized at specific places; particular places, for instance, are designated for such activities as working, worshiping, eating, or sleeping. Territorial boundaries delineate these places. These boundaries are defined by rules of property, which in any society structure the use and possession of scarce goods. In any society, moreover, there is a more or less regular division of labour. Yet another universal structural characteristic of human societies is the regulation of violence. The use of violence is everywhere a potentially disruptive force; at the same time, it is a means of coercion and coordination of activities. Human beings have formed political units, such as nations, within which the use of violence is strictly regulated and which, at the same time, are organized for the use of violence against outside groups.

In any society, furthermore, there are arrangements within the structure for sexual reproduction and the care and education of the young. These arrangements partly take the form of kinship and marriage relations. Finally, systems of symbolic communication, particularly language, everywhere structure the interactions between the members of a society.

Within the broad framework of these and other general features of human society, there is an enormous variety of social forms between and even within societies. Several theories have been developed to account for both the similarities and the varieties. In these theories certain aspects of social life are regarded as basic and, therefore, central components of the social structure.

Some social scientists use the concept of social structure as a device for creating an order for the various aspects of social life. Thus, the U.S. anthropologist George P. Murdock, in his *Social Structure* (1949), a comparative study of kinship systems, used the concept as a taxonomic scheme for classifying, comparing, and correlating aspects of kinship systems of different societies. In other studies, the concept is of greater theoretical importance; it is regarded as an explanatory concept, a key to the understanding of human social life. Some of the more prominent of these theories are reviewed here.

Murdock's taxonomic scheme

Structural functionalism. A.R. Radcliffe-Brown, a British social anthropologist, gave the concept of social structure a central place in his approach and connected it to the concept of function. In his view, the components of the social structure have indispensable functions for each other—the continued existence of the one component is dependent on that of the others—and for the society as a whole, which is seen as an integrated, organic entity. Radcliffe-Brown defined the social structure empirically as patterned, or "normal," social relations (those aspects of social activities that conform to accepted social rules or norms). These rules bind society's members to socially useful activities.

Structural functionalism was elaborated further by Talcott Parsons, a U.S. sociologist, who, like Radcliffe-Brown, was strongly influenced by the French social scientist Émile Durkheim. While Radcliffe-Brown focused on so-called primitive societies, Parsons attempted to formulate a theory that was valid for large and complex societies as well. For Parsons, the social structure is essentially normative; it consists of "institutionalized patterns of normative culture." Social behaviour is structured insofar as it conforms to norms, ranging from general ideas of right and wrong (values) to specific rules of behaviour in specific situations. These rules vary according to the positions of the individual actors: they define different roles, such as various occupational roles, or the roles of husband–father and wife–mother. Norms also vary according to the type of activities or sphere of life: they form clusters called social institutions, such as the institution of property or the institution of marriage. Norms, roles, and institutions are components of the social structure on different levels of complexity.

Relation of structure and norms

Theories of class and power. Parsons' work has been criticized for several reasons. One has been the comparatively meagre attention he paid to inequalities of power, wealth, and other social rewards. Other social theorists, including functionalists like the U.S. sociologist Robert K. Merton, have given these distributional properties a more central place in their concepts of social structure. For Merton and others, the social structure consists not only of normative patterns but also of the inequalities of power, status, and material privileges, which give the members of a society widely different opportunities and alternatives.

Social structure by class

In complex societies these inequalities define different strata, or classes, which form the stratification system, or class structure, of the society. Both aspects of the social structure, the normative and the distributive aspect, are strongly interconnected, as may be inferred from the observation that members of different classes often have different and even conflicting norms and values.

This leads to a consideration contrary to structural functionalism: certain norms in a society may be established, not because of any general consensus about their moral value, but because they are forced upon the population by those who have both the interest and the power to do so. To take one example, the "norms" of apartheid in South Africa reflect the interests and values of only one section of the population, which has the power to enforce them upon the majority. In theories of class and power this argument has been generalized: norms, values, and ideas are explained as the result of the power inequalities between groups with conflicting interests.

The most influential theory of this type has been Marxism, or historical materialism. The Marxian view is succinctly summarized in Marx's phrase that "the ideas of the ruling class are, in every age, the ruling ideas." These ideas are regarded as reflections of class interests and are connected to the power structure, which is identified with the class structure. This Marxian model, which was claimed to be particularly valid for capitalist societies, has met with several criticisms. One basic problem is its distinction between economic structure and spiritual superstructure, which are identified with social being and consciousness, respectively. This suggests that economic activities and relations are in themselves somehow not conscious, as if they were conceivable without knowing and thinking human beings.

The Marxian view of structure

Nevertheless, the Marxian model has become influential even among non-Marxist social scientists. The distinction between material structure and nonmaterial superstructure continues to be reflected in sociological textbooks as the distinction between social structure and culture. Social structure here refers to the ways people are interrelated or interdependent; culture refers to the ideas, knowledge, norms, customs, and capacities that they have learned and share as members of a society.

Structuralism. The concept of structure in the study called structuralism, as in structural functionalism and the class and power theories, is theoretical and explanatory. Unlike those other studies, however, it is not descriptive. The concept here refers to the underlying, unconscious regularities of human expressions, which are not observable but explain what is observed. Claude Lévi-Strauss, a French anthropologist, derived this concept from structural linguistics as developed by the Swiss linguist Ferdinand de Saussure. Any language is structured in the sense that its elements are interrelated in nonarbitrary, regular, rule-bound ways; a competent speaker of the language largely follows these rules without being aware of doing so. The task of the theorist is to detect this underlying structure, including the rules of transformation that connect the structure to the various observed expressions.

Lévi-Strauss and underlying structure

According to Lévi-Strauss, this same method can be applied to social and cultural life in general. He constructed theories concerning the underlying structure of kinship systems, myths, and customs of cooking and eating. The structural method, in short, purports to detect the common structure of widely different social and cultural forms. The structure does not determine the concrete expressions; the variety of expressions it generates is potentially unlimited. The structures that generate the varieties of social and cultural forms ultimately reflect, according to Lévi-Strauss, basic characteristics of the human mind.

Structuralism became an intellectual fashion in the 1960s in France, where such different writers as Roland Barthes, Michel Foucault, and Louis Althusser were also regarded as representatives of the new theoretical current. Structuralism in this wide sense, however, is not one coherent theoretical perspective. The Marxist structuralism of Althusser, for example, is far removed from Lévi-Strauss's anthropological structuralism. The structural method, when applied by different scholars, appears to lead to different results.

The criticisms launched against structural functionalism, class theories, and structuralism indicate that the concept of social structure is problematic. Yet the notion of social structure is not so easy to dispense with, because it expresses ideas of continuity, regularity, and interrelatedness in social life. Other terms are often used that have similar, but not identical, meanings, such as social network, social figuration, or social system. The British sociologist Anthony Giddens has suggested the term "structuration" in order to express the view that social life is, to a certain extent, both dynamic and ordered.

SOCIAL CHANGE

Social change in the broadest sense is any change in social relations. In this sense, social change is an ever-present phenomenon in any society. In order to give the concept a more restricted meaning, it has been defined as change

of the social structure. A distinction is made then between processes within the social structure, which serve, at least partially, to maintain the structure (social dynamics), and processes that modify the structure (social change). Because the concept of social structure does not have one generally accepted and unambiguous meaning, however, this distinction does not clearly determine which social processes belong to the field of social change.

The specific meaning of social change depends first of all on the social entity considered. Changes in a small group may be important on the level of that group itself, but negligible on the level of the larger society. Similarly, the observation of social change depends on the time span taken; most short-term changes are negligible if a social development is studied in the long run. Even if one abstracts from small-scale and short-term changes, social change is a general characteristic of human societies: customs and norms change, inventions are made and applied, environmental changes lead to new adaptations, conflicts result in redistributions of power.

Biological roots of change

This universal human potential for social change has a biological basis. It is rooted in the flexibility and adaptability of the human species—the near absence of biologically fixed action patterns on the one hand and the enormous capacity for learning, symbolizing, and creating on the other hand. The human biological constitution makes changes possible that are not biologically (genetically) determined. Social change, in other words, is only possible by virtue of biological characteristics of the human species, but the nature of the actual changes cannot be reduced to these species traits.

Historical background. Several ideas of social change have been developed in various cultures and historical periods. Three of them may be distinguished as the most basic: (1) the idea of decline or degeneration, or, in religious terms, the fall from an original state of grace; (2) the idea of cyclical change, a pattern of subsequent and recurring phases of growth and decline; and (3) the idea of continuous progress. These three ideas were already prominent in Greek and Roman antiquity and have characterized Western social thought from that time. The concept of progress, however, became the most influential idea, especially since the 18th-century Enlightenment. Social thinkers like Anne-Robert-Jacques Turgot and the Marquis de Condorcet in France and Adam Smith and John Millar in Scotland advanced theories on the progress of human knowledge and technology.

Social evolution as progress

Progress was the key idea in 19th-century theories of social evolution, and evolutionism was the common core shared by the most influential social theories of the century. Evolutionism implied that mankind as a whole progresses along one line of development; that this development is predetermined and inevitable, since it corresponds to definite laws; that some societies are more advanced in this development than other ones; and that Western society is the most advanced and therefore indicates the future of the rest of mankind. Auguste Comte, a French philosopher and sociologist, advanced a "law of three stages," according to which mankind progresses from a theological stage, which is dominated by religion, through a metaphysical stage, in which abstract speculative thinking is most prominent, and onward toward a positivist stage, in which scientific theories based on empirical research come to dominate.

The most encompassing theory of social evolution was developed by Herbert Spencer, who, unlike Comte, linked social evolution to biological evolution. According to Spencer, biological organisms and human societies follow the same universal, natural evolutionary law: "a change from a state of relatively indefinite, incoherent, homogeneity to a state of relatively definite, coherent, heterogeneity." In other words, as societies grow in size, they become more complex; their parts differentiate, specialize into different functions, and become, consequently, more interdependent.

Evolutionary thought also dominated the new field of social and cultural anthropology in the second half of the 19th century. Anthropologists such as Sir Edward Burnett Tylor and Lewis Henry Morgan classified contemporary societies on an evolutionary scale. Morgan ranked them from "savage" through "barbarian" to "civilized." Tylor postulated an evolution of religious ideas from animism through polytheism to monotheism. Morgan classified societies on the basis of the level of technology, or sources of subsistence, which he connected with the kinship system. He assumed that monogamy was preceded by polygamy, and patrilineal descent by matrilineal descent.

Morgan's evolutionary rankings

Marx and Friedrich Engels too were highly influenced by evolutionary ideas. The Marxian distinctions between primitive communism, the Asiatic mode of production, ancient slavery, feudalism, capitalism, and future socialism may be interpreted as a list of stages in one evolutionary development, although the Asiatic mode did not fit well in this scheme. Marx and Engels were impressed by Morgan's anthropological theory of evolution, which became evident in Engels' book *Der Ursprung der Familie, des Privateigentums und des Staats* (1884; *The Origin of the Family, Private Property and the State*).

The originality of the Marxian theory of social development lay in its combination of dialectics and gradualism. In Marx's view social development was a dialectical process: the transition from one stage to another took place through a revolutionary transformation, which was preceded by increasing deterioration of society and intensifying class struggles. Underlying this discontinuous development was the more gradual development of the forces of production (technology and organization of labour).

Marx was influenced by the countercurrent of Romanticism, which was opposed to the idea of progress. This influence was evident in his notion of "alienation," which meant that in the course of social development people had increasingly lost control over the social forces that they had produced by their own activities. Romantic counterprogressivism was, however, much stronger in the work of other social theorists of the century, such as Ferdinand Tönnies, a German sociologist. He distinguished between the community (*Gemeinschaft*), in which people were bound together by common traditions and ties of affection and solidarity, and the society (*Gesellschaft*), in which social relations had become contractual, rational, and nonemotional.

Gemeinschaft and *Gesellschaft*

Durkheim and Max Weber, sociologists who began their careers at the end of the 19th century, showed ambivalence toward the ideas of progress. Durkheim regarded the increasing division of labour as a basic process, which was at the roots of modern individualism, but could also lead to "anomie," or lack of moral norms. Weber rejected evolutionism by arguing that the development of Western society was quite different from that of other civilizations and therefore historically unique. It was characterized, according to Weber, by a peculiar type of rationalization, which had brought modern capitalism, modern science, and rational law, but also, on the negative side, a "disenchantment of the world" and increasing bureaucratization.

The work of Durkheim, Weber, and other social theorists around the turn of the century marked a transition from evolutionism toward more static theories. Evolutionary theories were criticized on empirical grounds—they could be refuted by a growing mass of research findings—and because of their determinism and Western-centred optimism. Theories of cyclical change that denied long-term progress gained popularity in the first half of this century; these included the theory of the Italian economist and sociologist Vilfredo Pareto on the "circulation of elites" and those of Oswald Spengler and Arnold Toynbee on the life cycle of civilizations. Although the interest in long-term social change never disappeared, it faded to the background, especially when, from the 1920s until the 1950s, functionalism, emphasizing an interdependent social system, became the dominant paradigm both in anthropology and in sociology. "Social evolution" was substituted for the more general and neutral concept of "social change."

Decline of evolutionism

From the 1950s and increasingly through the 1960s and 1970s there was a revival of interest in long-term social change. Neo-evolutionist theories were proclaimed by several anthropologists, including Ralph Linton, Leslie A. White, Julian H. Steward, Marshall D. Sahlins, and Elman Rogers Service. These authors hold to the idea of social

evolution as a long-term development of mankind, which is patterned and cumulative in some respects. Neo-evolutionism differs from 19th-century evolutionism in that it does not assume that all societies go through the same stages of development; much attention is paid to variations between societies as well as to relations of influence among them. The latter concept has come to be known by the term acculturation. Moreover, social evolution is not regarded as predetermined or inevitable but is rather conceived in terms of probabilities. Finally, evolutionary development is not equated with progress.

Modernization theories

The revival of interest in long-term social change has been partly induced by the problems of the so-called underdeveloped countries. In order to explain the gaps between rich and poor countries, Western sociologists and economists in the 1950s and 1960s elaborated modernization theories. These theories implied a covertly Western-centred evolutionism insofar as they assumed that poor countries had stagnated on a relatively low level of development and could and should develop, or modernize, in the direction of a Western-type society. Modernization theories have been criticized for their lack of attention to international power relations, in which the richer countries dominate the poorer ones. These relations have been brought into the centre of attention by more recent theories of international dependency or, in Immanuel Wallerstein's terms, the "world capitalist system."

Since about 1965 there has been some convergence between sociology and anthropology on the one hand and history on the other. Historians have become interested in theories of long-term social change, while many sociologists and anthropologists increasingly turn toward history for the empirical testing and refinements of their theoretical viewpoints.

Patterns of social change. The common assumption of theories of social change, old and new, is that the course of such change is not arbitrary but, to a certain degree, regular or patterned. The three traditional ideas of social change—those of decline, cyclical change, and progress—have influenced modern theories. However, insofar as these theories are nonnormative, or scientifically determined, they do not distinguish explicitly between decline and progress. Such values cannot be derived from empirical observations alone but depend on normative evaluations, or value judgments. In nonnormative terms, then, two basic patterns of social change emerge: the cyclical and the one-directional. Often the time span of the change determines which pattern is observed.

Cyclical change. A regular alternation of stages characterizes cyclical change. Much of ordinary social life is organized in cyclical changes: those of the day, the week, and the year. These short-term cyclical changes may be regarded as conditions necessary to structural stability. Other changes that have a more or less cyclical pattern are less regular. For example, business cycles, recurrent phenomena of capitalist and industrial societies, are patterned to some extent yet hard to predict in concrete cases. A well-known theory of the business cycle is that of the Soviet economist Nikolay D. Kondratyev, who tried to show the recurrence of long waves of economic boom and recession on an international scale. He charted the waves from the end of the 18th century, with each complete wave comprising a period of about 50 years. Subsequent research has shown, however, that the patterns in different countries have been far from identical.

Kondratyev's waves

Long-term cyclical changes are rendered by theories on the birth, growth, flourishing, decline, and death of civilizations. Toynbee conceived world history in this way in the first volumes of *A Study of History* (1934–61), as did Spengler in his *Untergang des Abendlandes* (1918–22; *Decline of the West*). These theories have been criticized for their conception of civilizations as natural entities with sharp boundaries because this tends to neglect the interrelations between civilizations.

One-directional change. A continuation in terms of more or less characterizes one-directional change. Such change is usually cumulative; it implies growth or increase, such as that of population density, the size of organizations, or the level of production. However, the direction of the change may also be one of decrease, or a combination of growth and decrease. An example of this last process is what the American cultural anthropologist Clifford Geertz has called "involution," found in some agrarian societies: population growth coupled with decreasing per capita wealth. Or the change may be a shift from one to the other pole of a continuum—from religious to scientific ways of thinking, for example. Such a change may be defined as either growth (of scientific knowledge) or decline (of religion).

The concept of involution

The simplest type of one-directional change is linear: the extent of social change is constant over time. Another type of regular social change is exponential growth, in which the growth percentage is constant over time and the change accelerates correspondingly. Population growth and production growth often approximate this pattern during some periods of time.

A pattern of long-term growth may also conform to a three-stage S-curve: at the beginning of the period under consideration the change is almost imperceptibly slow, then accelerates, then slackens, until it approaches a supposed upper limit. The model of the demographic transition in industrializing countries exhibits this pattern. In the first stage, premodern or preindustrial, both the birthrate and the mortality rate are high, and, consequently, the population grows very slowly; then mortality decreases, and the population grows much faster; in the third stage both the birthrate and the mortality rate have become low, and the population growth approaches zero. The same model has been suggested, more hypothetically, for the rate of technological and scientific change.

Combined patterns of change. Cyclical and one-directional changes may be combined in one way or another. Very often short-term changes are cyclical while long-term development is in one direction. Figures on the production rates of industrializing countries conform, more or less, to this pattern, short-term business cycles occurring within long-term economic growth.

All of these pattern models cannot be applied simply and easily to social reality. They are at best approximations of parts of social reality. Comparing the model with the reality is not always possible because of a lack of reliable data. Moreover, and more importantly, many social processes do not lend themselves to precise quantitative measurement. Processes like bureaucratization or secularization, for example, can be defined as changes in a certain direction, but it is hard to measure the extent to which the change in a given period has taken place. It is doubtful that models like those of linear or exponential growth can be used in such cases.

Determination of long-term change

It remains to be seen whether or not long-term social change in a certain direction may be ascertained. Many investigations have sought answers to this question for Western society since the Middle Ages. The transformation of medieval society into the Western nations of the 20th century may be conceived in terms of several interconnected, long-term, one-directional changes; some of the more important of these include commercialization, increasing division of labour, growth of production, formation of national states, bureaucratization, growth of technology and science, secularization, urbanization, spread of literacy, increasing social and geographical mobility, and growth of organizations. Many of these changes have also occurred in non-Western societies. Most changes have not originated in the West, but some important changes did originate there—particularly such complex transformations as the rise of capitalism and the Industrial Revolution. These subsequently had a strong impact on non-Western societies. Groups of people outside western Europe have been incorporated in a global division of labour, in which the Western nation-states dominated both politically and economically.

The extent to which these changes are part of a global, long-term social development is the central question of social evolution, which is conceived as a very long-term one-directional change for mankind as a whole. Although knowledge concerning this question is far from complete, some very broad and general trends may be hypothesized on firm ground. First, technological innovations and

General trends of social evolution

growing empirical knowledge led to an increasing control of natural forces for the satisfaction of human needs. Parts of this development were the use and control of fire, the cultivation of plants and the domestication of animals (dating from about 8000 BC), the use of metals, and the process of industrialization. This technological development, combined with long-term capital accumulation, led to rising production levels and, therefore, made possible population growth and increasing population density. Energy production and consumption grew, if not per capita then at least per square mile.

Interconnected with technological development and growth of production were the process of division of labour and social differentiation. On the one hand, it was only by division of labour and corresponding specialization of knowledge and abilities that the technical control of natural forces could increase beyond certain limits. On the other hand, the growth of production as a result of technological innovations contributed to further social differentiation; more people, in other words, could specialize in activities that were not immediately necessary for survival. Growing size and density of populations and social differentiation led to increased interdependence between growing numbers of people over longer distances. In hunting and gathering societies people were strongly interdependent within their small bands, depending on very little outside their groups. In modern times, most of the world's people are enmeshed in worldwide networks of interdependence.

These processes were not inevitable in the sense that they corresponded to any "law" of social change. They had the tendency, however, to spread whenever they occurred. For example, once the set of transformations known as the agrarian revolution had taken place anywhere in the world, their extension over the rest of the world was predictable. Societies that adopted these innovations grew in size and became more powerful. As a consequence, other societies had only three options: to be conquered and incorporated by a more powerful agrarian society; to adopt the innovations; or to be driven away to marginal places of the globe. Something similar might be said of the Industrial Revolution and other power-enhancing innovations, such as bureaucratization and the introduction of more destructive weapons. This last example illustrates that these processes should not be equated with progress in general.

Causal connections between social processes

Explanations of social change. One way of explaining social change is to show causal connections between two or more processes. This may take the form of a kind of determinism or reductionism, which explains all social change by reducing it to one supposed autonomous and all-determining causal process. A more cautious assumption is that one process has relative causal priority, without implying that this process is completely autonomous and all-determining. Following are some of the processes conceived as having caused social change.

Natural environment. Changes in the natural environment may vary from climatic ones to those caused by the spread of diseases. For example, both the worsening of climatic conditions and the epidemics of the Black Death have been submitted as factors that explain the crisis of feudalism in 14th-century Europe. Changes in the natural environment may be either independent of human social activities or the result of these activities. Deforestation and erosion, air pollution, and the exhaustion of natural resources belong to the last category, and they in turn may have far-reaching social consequences.

Demographic processes. Population growth and increasing population density represent, in particular, demographic forms of social change. Population growth may lead to geographical expansion of a society, military conflicts, and the intermingling of cultures. Increasing population density may also stimulate technological innovations, which may increase division of labour and social differentiation, commercialization, and urbanization. This has been observed as affecting western Europe from the 11th to the 13th century and England in the 18th century, where population growth was a factor in the Industrial Revolution. On the other hand, population growth may contribute to economic stagnation and increasing poverty, as may be witnessed in several Third World countries today.

Technological innovations. According to several theories of social evolution, technological innovations are regarded as the most important determinants of societal change. The social significance of such technological breakthroughs as the invention of the smelting of iron, the introduction of the plow in agriculture, the invention of the steam engine, and the development of the computer is indeed evident. Of course, it is possible to dispute the relative importance of such innovations when compared to other determinants of social change.

Economic processes. Technological changes are often considered in conjunction with economic processes, including the formation and extension of markets, modifications of property relations (such as the change from feudal lord–peasant relations to contractual proprietor–tenant relations), and changes in the organization of labour (such as the change from independent craftsmen to factories). Historical materialism, as developed by Marx and Engels, is the most influential theory that gives priority to economic processes, but it is not the only one. Materialist theories have been developed even in opposition to Marxism, one being the "logic of industrialization" thesis by the U.S. scholar Clark Kerr, which states that industrialization everywhere has similar consequences, whether the property relations are called capitalist or communist.

Ideas. Other theories have stressed the significance of ideas in the causation of social change. Comte's law of three stages is such a theory. Weber regarded religious ideas as important in contributing to economic development or stagnation; according to his controversial thesis, the individualistic ethic of Christianity, and in particular Protestantism, partially explains the rise of the capitalist spirit, which brought economic dynamism in the West.

Social movement. A change of collective ideas is not merely an intellectual process; it is often connected to the formation of a new social movement. This in itself might be regarded as a potential cause of social change. Weber called attention to this factor in conjunction with his concept of "charismatic leadership." The charismatic leader, by virtue of the extraordinary personal qualities attributed to him, is able to create a group of followers who are willing to break established rules.

Political processes. Changes in the regulation of violence, in the nature of the state organization, and in international relations may also determine social change. For example, the German sociologist Norbert Elias has analyzed the formation of states in western Europe as a relatively autonomous process that led to an increasing control of violence and, consequently, to rising standards of self-control. According to recent theories of political revolution, the functioning of the state apparatus itself and the nature of interstate relations are of decisive importance in the outbreak of a revolution: it is only when the state is not able to fulfill its basic functions of maintaining law and order and territorial integrity that revolutionary groups have any chance of success.

Criticism of determinism

Each of these processes is a possible determinant of other ones; none of them is the only determinant. One reason why deterministic, or reductionist, theories run into difficulties is that the process they use to explain the process as a whole is actually not autonomous but has to be itself explained. Moreover, social processes are often intertwined to such a degree that it would be misleading to consider them separately. For example, there are no sharp and fixed borderlines between economic and political processes, nor between economic and technological processes. Technological change may in itself be regarded as a specific type of cultural or conceptual change. The causal connections between distinguishable social processes are a matter of degree and vary over time.

Mechanisms of social change. The scope of any causal explanation of social change in which initial conditions or basic processes are specified is limited. A more general and theoretical way of explaining is to construct a model of recurring mechanisms of social change. Such mechanisms, incorporated in different theoretical models, include the following.

Mechanisms of one-directional change: accumulation, selection, and differentiation. Some evolutionary theories stress the essentially cumulative nature of human knowledge. Because human beings are innovative, they add to existing knowledge, replacing less adequate ideas and practices with more adequate ones. As they learn from mistakes, they select new ideas and practices in a trial-and-error process (sometimes compared to the process of natural selection). The expansion of collective knowledge and capabilities beyond a certain limit is only possible by specialization and differentiation. Growth of technical knowledge stimulates capital accumulation, which leads to rising production levels. Population growth may also be incorporated in this model of cumulative evolution: it is by the accumulation of collective technical knowledge and means of production that human beings can multiply their numbers; this growth then leads to new problems that stimulate further innovations.

Mechanisms of curvilinear and cyclical change: saturation and exhaustion. Models of one-directional change assume that change in a certain direction induces further change in the same direction; models of curvilinear or cyclical change, on the other hand, assume that change in a certain direction creates the conditions for change in another (perhaps even the opposite) direction. More specifically, it is often assumed that growth has its limits and that in approaching these limits the change curve will inevitably be bent. Ecological conditions like the availability of natural resources, in particular, set limits to population growth and economic growth.

Shorter term cyclical changes are explained by comparable mechanisms. Some theories of the business cycle, for example, assume that the economy is saturated periodically with capital goods: investments become less necessary and less profitable, the rate of investments diminishes, and a negative spiral resulting in a recession sets in. After a period of time, however, essential capital goods will have to be replaced: investments are pushed up again and a phase of economic expansion begins.

Group conflict as a cause of change

Conflict, competition, and cooperation. Group conflict has often been viewed as a basic mechanism of social change, especially of those radical and sudden social transformations identified as revolutions. Marxists in particular tend to depict social life in capitalist society as a struggle between a ruling class, which wishes to maintain the system, and a dominated class striving for radical change; social change then is the result of that struggle. These ideas are basic to what Dahrendorf has called a conflict model of society.

The notion of conflict becomes more relevant for the explanation of social change if it is broadened to include competition between rival groups as well. Nations, firms, universities, sports associations, and artistic schools are groups between which such rivalry occurs. Competition stimulates the introduction and diffusion of innovations, especially when they are potentially power-enhancing. Thus, the leaders of non-Western states feel the necessity of adopting Western science and technology, even though their ideology may be anti-Western, because it is only by these means that they can maintain or enhance national autonomy and power.

Additionally, competition may lead to the growing size and complexity of the entities involved. The classic example of this process, analyzed by Marx, is the tendency in capitalism for monopolies to form as small firms are driven out of competition by larger ones. Marx's analysis has been applied to another area by Norbert Elias, who explained the formation of national states in western Europe as the result of competitive struggles between feudal lords.

Competition is also put forward in individualistic theories, which conceive social change as the result of the actions of individuals pursuing their self-interest. With the help of game theory and other mathematical devices it has been shown that individuals acting on the basis of self-interest will cooperate, given certain conditions, in widening social networks.

Tension and adaptation. In structural functionalism, social change is regarded as the adaptive response to some tension within the social system. When some part of an integrated social system changes, a tension between this and other parts of the system is created, which will be resolved by the adaptive change of the other parts. An example is what the U.S. sociologist William Fielding Ogburn has called cultural lag, which refers in particular to a gap that develops between fast-changing technology and other slower paced sociocultural traits.

Cultural lag

Diffusion of innovations. Some social changes are to be regarded as the result of the diffusion of innovations, such as technological inventions, new scientific knowledge, new beliefs, or a new fashion in the sphere of leisure. Diffusion is not automatic but selective; an innovation is only adopted by people if they are motivated to do so and if it is compatible with important aspects of their culture. One reason for the adoption of innovations by larger groups is the example of higher status groups, which are reference groups for other people. Successful innovations, which affect the majority of the people of a society, tend to follow a pattern of diffusion from higher to lower status groups. More specifically, most early adopters of innovations in modern Western societies, according to several studies, are young, urban, and highly educated, with a high occupational status. Often they are motivated by the wish to distinguish themselves from the mass of the population. After diffusion has taken place, however, the innovation is no longer a symbol of distinction, which motivates the same group to look for something new again. This mechanism may explain the succession of trends in several fields.

Young urban professionals

Planning and institutionalization of change. Social change may be, to a certain extent, the result of goal-directed, large-scale social planning. The possibilities of planning by government bureaucracies and other large organizations have increased in modern societies. Most social planning is short-term, however; the goals of planning are often not attained, and, even if the planning is successful in terms of the stated goals, it often has unforeseen consequences. The wider the scope and the longer the time span of planning, the more difficult it is to attain the goals and to avoid unforeseen and undesired consequences. This has become especially clear in Communist societies, where the most serious efforts have been taken to put the ideal of integral and long-term planning into practice. Large-scale and long-term social developments in any society are still largely unplanned.

Planning implies institutionalization of change, but institutionalization does not imply planning. Many unplanned social changes in modern societies are institutionalized; they originate in organizations permanently oriented to innovation, such as universities and the research departments of governments and private firms, but their social repercussions are not controlled. It is in the fields of science and technology especially that change is institutionalized, producing social change that is partly intended and partly unintended.

These mechanisms of social change are not mutually exclusive. On the contrary, some of them are clearly interconnected. For example, innovation by specialized organizations is stimulated by competition. Several mechanisms may be combined in one explanatory model of social change.

CONCLUSION

The interconnection of structure and change

Social structure and social change are central theoretical concepts of the social sciences that refer to basic and complementary characteristics of social life in general—permanence, continuity, and repetitiveness on the one hand, dynamics and changeability on the other. Both concepts are interconnected: the social structure cannot be conceptualized adequately without some notion of actual or potential change, and social change as a more or less regular process is inconceivable without the notion of continuity. To the degree that change processes are regular and interconnected, social change itself is structured. Any separation of the two concepts, as though they refer to divergent fields, is therefore misleading. This is not to deny that the relative stress on either structural continuity or dynamic change varies in social scientific theories and empirical studies. Since about 1965 there has been a shift from "structure" to "change" in social theory. Change on

different levels—social dynamics in everyday life, short-term transformations and long-term developments in society at large—has become the focus of attention.

BIBLIOGRAPHY. A general reader on social structure is PETER M. BLAU (ed.), *Approaches to the Study of Social Structure* (1975). The most important theoretical works in structural functionalism are A.R. RADCLIFFE-BROWN, *Structure and Function in Primitive Society* (1952, reissued 1965, reprinted 1968); and TALCOTT PARSONS, *The Social System* (1951, reprinted 1964). For coverage of the debate on structural functionalism, see N.J. DEMERATH and RICHARD A. PETERSON (eds.), *System, Change, and Conflict* (1967, reprinted 1968). A more empirical type of functionalism is represented by ROBERT K. MERTON, *Social Theory and Social Structure: Toward the Codification of Theory and Research*, new ed. (1968), in which due consideration is given the distributive aspects of the social structure. These are stressed even more by PETER M. BLAU, *Inequality and Heterogeneity: A Primitive Theory of Social Structure* (1977). RALF DAHRENDORF, *Class and Class Conflict in Industrial Society* (1959; originally published in German, 1957), advances a power-and-conflict model of society. Other, more sophisticated power models are contained in PETER M. BLAU, *Exchange and Power in Social Life* (1964); STEVEN LUKES, *Power: A Radical View* (1974); and NORBERT ELIAS, *What Is Sociology?* (1978; originally published in German, 3rd ed., 1978). An introduction to structuralism is DAVID ROBEY (ed.), *Structuralism* (1973). CLAUDE LÉVI-STRAUSS, *Structural Anthropology*, 2 vol. (1963–76; originally published in French, 1958–73), contains several articles on the structural method and its applications. Examples of different empirical applications of the concept of social structure are GEORGE PETER MURDOCK, *Social Structure* (1949, reissued 1965); PETER M. BLAU and OTIS DUDLEY DUNCAN, *The American Occupational Structure* (1967, reprinted 1978); and PETER V. MARSDEN and NAN LIN (eds.), *Social Structure and Network Analysis* (1982). A synthesis of different views is offered by ANTHONY GIDDENS, *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis* (1979, reprinted 1983).

On the history of ideas concerning social change, see ROBERT A. NISBET, *Social Change and History: Aspects of the Western Theory of Development* (1969). An introduction to 18th- and 19th-century evolutionism is LOUIS SCHNEIDER, *Classical Theories of Social Change* (1967). Original texts in social evolutionism are HERBERT SPENCER, *The Principles of Sociology*, 3 vol. in 4 (1876–96, reprinted in 3 vol., 1975), and *Herbert Spencer: Structure, Function, and Evolution*, ed. by STANISLAV ANDRESKI (1971, reissued 1972); LEWIS HENRY MORGAN, *Ancient Society* (1877, reissued 1985); and EDWARD B. TYLOR, *Primitive Culture*, 2 vol. (1871, reissued 1970). Good selections of Marxian texts are KARL MARX, *Selected Writings in Sociology and Social Philosophy*, ed. by T.B. BOTTOMORE and MAXIMILIAN RUBEL (1956, reprinted 1964); and KARL MARX and FREDERICK ENGELS, *Selected Works*, 2 vol. (1935, reissued in 1 vol., 1968). The most influential study in Marxist evolutionism is FREDERICK ENGELS, *The Origin of the Family, Private Property and the State* (1902, reissued 1978; originally published in German, 1884). A criticism of Spencer's evolutionism is contained in ÉMILE DURKHEIM, *Émile Durkheim on the Division of Labor in Society* (1933, reissued 1984 as *The Division of Labor in Society;* originally published in French, 1893); while MAX WEBER, *The Protestant Ethic and the Spirit of Capitalism* (1930, reprinted 1985; originally published in German, 1920, in vol. 1 of his *Gesammelte Aufsätze*), contains a criticism of historical materialism.

Anthropological neo-evolutionism is represented by: LESLIE A. WHITE, *The Evolution of Culture: The Development of Civilization to the Fall of Rome* (1959); JULIAN H. STEWARD, *Theory of Culture Change: The Methodology of Multilinear Evolution* (1955, reprinted 1973); MARSHALL D. SAHLINS and ELMAN R. SERVICE (eds.), *Evolution and Culture* (1960, reprinted 1982); W.F. WERTHEIM, *Evolutie en revolutie: De golfslag der emancipatie* (1971), from which an abridged English trans., *Evolution and Revolution: The Rising Waves of Emancipation* (1974), was made; and ELMAN R. SERVICE, *Cultural Evolutionism: Theory in Practice* (1971). A sociological textbook with an evolutionary approach is GERHARD LENSKI and JEAN LENSKI, *Human Societies: An Introduction to Macrosociology*, 4th ed. (1982). S.N. EISENSTADT, *Tradition, Change, and Modernity* (1973, reprinted 1983), represents a sophisticated version of the modernization theory. Good examples of historical sociology are NORBERT ELIAS, *The Civilizing Process*, 2 vol. (1978; originally published in German, 1939); BARRINGTON MOORE, JR., *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World* (1966, reissued 1984); and IMMANUEL WALLERSTEIN, *The Modern World-System*, vol. 1, *Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century* (1974). Akin to these books are comprehensive historical studies on long-term developments, such as WILLIAM H. MCNEILL, *The Rise of the West: A History of the Human Community* (1963, reissued 1965); and FERNAND BRAUDEL, *Capitalism and Material Life, 1400–1800* (1974; originally published in French, 1967). An overview of this field is given by THEDA SKOCPOL (ed.), *Vision and Method in Historical Sociology* (1984).

General theoretical books on social change are: WILBERT E. MOORE, *Social Change*, 2nd ed. (1974), and *Order and Change: Essays in Comparative Sociology* (1967), two treatises in the functionalist tradition; EVA ETZIONI-HALEVY and AMITAI ETZIONI (eds.), *Social Change: Sources, Patterns, and Consequences*, 2nd ed. (1973), a reader representing various approaches; WILLIAM FIELDING OGBURN, *Social Change with Respect to Culture and Original Nature*, new ed. (1950, reprinted 1965); AMITAI ETZIONI, *The Active Society: A Theory of Societal and Political Processes* (1968, reprinted 1971), which explores the possibilities of planned change; ROBERT L. HAMBLIN, R. BROOKE JACOBSEN, and JERRY L.L. MILLER, *A Mathematical Theory of Social Change* (1973); HENRY TEUNE and ZDRAVKO MLINAR, *The Developmental Logic of Social Systems* (1978); and KENNETH E. BOULDING, *Ecodynamics: A New Theory of Societal Evolution* (1978, reprinted 1981).

Controversial theories on the cyclical development of civilizations have been advanced by OSWALD SPENGLER, *The Decline of the West*, 2 vol. (1922, reissued 1981–83; originally published in German, 1918–22); and ARNOLD J. TOYNBEE, *A Study of History*, 12 vol. (1934–61, reprinted 1948–61). A theory of the circulation of elites can be found in VILFREDO PARETO, *The Mind and Society: Treatise on General Sociology*, 4 vol. (1935, reprinted 1983; originally published in Italian, 2nd ed., 3 vol., 1923), and *Sociological Writings*, ed. by S.E. FINER (1966, reprinted 1976). An empirical test of theories of economic growth and the business cycle is given by ANGUS MADDISON, *Phases of Capitalist Development* (1982). The concept of involution is explained by CLIFFORD GEERTZ, *Agricultural Involution: The Process of Ecological Change in Indonesia* (1963, reprinted 1968).

The significance of demographic processes is analyzed by CARLO M. CIPOLLA, *The Economic History of World Population*, 7th ed. (1978), and an analysis of one stage of development is presented in MARK NATHAN COHEN, *The Food Crisis in Prehistory: Overpopulation and the Origins of Agriculture* (1977, reprinted 1979). A classic account of the influence of technological change is V. GORDON CHILDE, *Man Makes Himself*, rev. ed. (1951, reissued 1983). CLARK KERR *et al.*, *Industrialism and Industrial Man: The Problems of Labor and Management in Economic Growth*, 2nd ed. (1964, reissued 1973), represents a non-Marxist materialist view. Theories of political revolution are developed by CRANE BRINTON, *The Anatomy of Revolution*, rev. and expanded ed. (1965); and THEDA SKOCPOL, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China* (1979). EVERETT M. ROGERS, *Diffusion of Innovations*, 3rd ed. (1983), deals with the social aspects of technological innovations. The individualistic approach to social change processes is exemplified by DOUGLASS C. NORTH, *Structure and Change in Economic History* (1981); MANCUR OLSON, *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities* (1982, reprinted 1984); and ROBERT AXELROD, *The Evolution of Cooperation* (1984).

An influential criticism of deterministic theories of social development is KARL R. POPPER, *The Poverty of Historicism*, 2nd ed. (1960). Examples of pessimistic social forecasting with much attention to ecological conditions are DONELLA H. MEADOWS, *et al.*, *The Limits to Growth*, 2nd ed. (1974, reprinted 1982); and MIHAJLO MESAROVIC and EDUARD PESTEL, *Mankind at the Turning Point* (1974, reissued 1976). Much more optimistic examples are HERMAN KAHN and ANTHONY J. WIENER, *The Year 2000* (1967); DANIEL BELL, *The Coming of Post-Industrial Society: A Venture in Social Forecasting* (1973, reprinted 1976); and CLARK KERR, *The Future of Industrial Societies: Convergence or Continuing Diversity?* (1983). For a history of ideas about future social developments, see KRISHAN KUMAR, *Prophecy and Progress: The Sociology of Industrial and Post-Industrial Society* (1978). (N.Wi.)

Social Welfare

The basic concerns of social welfare—poverty, disability and disease, the dependent young and elderly—are as old as society itself. The laws of survival once severely limited the means by which these concerns could be addressed; to share another's burden meant to weaken one's own standing in the fierce struggle of daily existence. As societies developed, however, with their patterns of dependence between members, there arose more systematic responses to the factors that rendered individuals, and thus society at large, vulnerable.

Religion and philosophy have tended to provide frameworks for the conduct of social welfare. The edicts of the Buddhist emperor Aśoka in India, the sociopolitical doctrines of ancient Greece and Rome, and the simple rules of the early Christian communities are a few examples of systems that addressed social needs. The Elizabethan Poor Laws in England, which sought relief of paupers through care services and workhouses administered at the parish level, provided precedents for many modern legislative responses to poverty. In Victorian times a more stringent legal view of poverty as a moral failing was met with the rise of humanitarianism and a proliferation of social reformers. The social charities and philanthropic societies founded by these pioneers formed the basis for many of today's welfare services.

Because perceived needs and the ability to address them determine each society's range of welfare services, there exists no universal vocabulary of social welfare. In some countries a distinction is drawn between "social services," denoting programs, such as health care and education, that serve the general population, and "welfare services," denoting aid directed to vulnerable groups, such as the poor, the disabled, or the delinquent. According to another classification, remedial services address the basic needs of individuals in acute or chronic distress; preventive services seek to reduce the pressures and obstacles that cause such distress; and supportive services attempt, through educational, health, employment, and other programs, to maintain and improve the functioning of individuals in society. Social welfare services originated as emergency measures that were to be applied when all else failed. However, they are now generally regarded as a necessary function in any society and a means not only of rescuing the endangered but also of fostering a society's ongoing, corporate well-being.

This article treats first the personal social services—those provided on an individual basis to persons in need; collectively these services constitute the professional field of social work. A discussion of government-sponsored social security programs, such as social assistance and social insurance, follows.

The article is divided into the following sections:

Social work: the personal social services 372
- Modern evolution 372
- Major areas of concern 373
  - Family welfare
  - Child welfare
  - Youth welfare
  - Welfare of the elderly
  - Group welfare
  - Welfare of the sick and disabled
  - Welfare of the mentally ill
- The work of the personal social services 375
  - Social work training
  - Administration of services
- Conclusion 378

Social security: government welfare programs 378
- The rationale for social security 379
- Historical evolution 379
- Methods of provision 381
  - Legal liability
  - Provident schemes
  - Social insurance
  - Benefits to all residents
  - Social assistance
  - Negative income tax
- Cash benefit programs 382
  - Pensions
  - Disability and sickness benefits
  - Unemployment benefits
  - Family, maternity, and parental allowances
  - Benefits for survivors and single parents below pension age
  - Variations in provision between countries
- Benefits in kind 386
- Administration and finance 388
  - Administration of social security
  - Financing of social security
  - The rising cost of social security
- Criticisms 391

Bibliography 392

Social work: the personal social services

The majority of personal social services are rendered on an individual basis to people who are unable, whether temporarily or permanently, to cope with the problems of everyday living. Recipients include families faced with loss of income, desertion, or illness; children and youths whose physical or moral welfare is at risk; the sick; the disabled; the frail elderly; and the unemployed. When possible, services are also directed toward preventing threats to personal or family independence.

Social services generally place a high value on keeping families together in their local communities, organizing support from friends or neighbours when kinship ties are weak. Where necessary, the services provide substitute forms of home life or residential care, and play a key role in the care and control of juvenile delinquents and other socially deviant groups, such as drug and alcohol abusers.

MODERN EVOLUTION

In the advanced industrial societies the personal social services have always constituted a "mixed economy of welfare," involving the statutory, voluntary, and private sectors of welfare provision. Although the role of personal social services is crucial, they account for only a small proportion of total welfare expenditures. The most substantial increases in expenditures have occurred in social security systems, which provide assistance to specific categories of claimants on the basis of both universal and selective criteria. The development of modern social security systems from the 1880s reflects not only a gradual but fundamental change in the aims and scope of social policy but also a dramatic shift in expert and popular opinion with regard to the relative significance of the social and personal causes of need (see below *Social security: government welfare programs*).

Poor relief

In the belief that personal shortcomings were the chief cause of poverty and of people's inability to cope with it, the major 19th-century systems of poor relief in western Europe and North America tended to withhold relief from all but the truly destitute, to whom it was given as a last resort. This policy was intended as a general deterrent to idleness. The poor-law relieving officer was the precursor of both the public assistance officials and the social work-

ers of today in his command of statutory financial aid. The voluntary charitable agencies of the time differed on the relative merits of deterrent poor-law services on the one hand, implying resistance to the growth of statutory welfare, and on the provision of alternative assistance to the needy, coupled with the extension of statutory services, on the other hand. From the 1870s the Charity Organization Society and similar bodies in the United States, Britain, and elsewhere held strongly to the former option, and their influence was widespread until the outbreak of World War II.

Settlement movement

The settlement movement in Britain and the United States drew voluntary workers into direct contact with the serious material disadvantages suffered by the poor. The pioneer of this movement was the vicar Samuel A. Barnett, who in 1884 with his wife and a number of university students "settled" in a deprived area of London, calling their neighbourhood house Toynbee Hall. Two visitors to this settlement soon introduced the movement into the United States—Stanton Coit, who founded Neighborhood Guild (later University Settlement) on the Lower East Side of New York City in 1886, and Jane Addams, who with Ellen Gates Starr founded Hull House on the Near West Side of Chicago in 1889. From these prototypes the movement spread to other U.S. cities and abroad through Europe and Asia.

The origins of modern social casework can be traced to the appointment of the first medical almoners in Britain in the 1880s, a practice quickly adopted in North American and most western European countries. The almoners originally performed three main functions: ascertaining the financial eligibility and resources of patients faced with the rising costs of medical care, providing counseling services to support patients and their families during periods of ill health and bereavement, and procuring adequate practical aids and other forms of home care for discharged patients. Elsewhere secular and religious charitable associations providing financial help, educational welfare, and housing for the poor began to employ social workers.

By the turn of the century there were various schemes for organizing charitable work on "scientific" principles according to nationally agreed standards of procedure and services. In Britain, the United States, Germany, and, later, Japan, leading charities worked in conjunction with poor-law and public assistance authorities, an approach endorsed in 1909 in the majority report of the British Royal Commission on the Poor Law. The first schools of social work, usually run by the voluntary charitable agencies, appeared in the 1890s and early 1900s in London, New York City, and Amsterdam, and by the 1920s there were similar ventures in other parts of western Europe and North America and in South America. The training programs combined casework methods and other practical forms of intervention and support, with particular emphasis on working in cooperation with individuals and families to restore a level of independence.

Early social work training

From the 1900s onward the surveys conducted by Charles Booth in London and Seebohm Rowntree in York and by other researchers began to transform conventional views of the role of the state in social welfare and the relief of poverty, and the social causes of poverty came under scrutiny. At the same time, the scope of social work was growing, with the spread of settlement houses, to include group work and community action.

In most countries social welfare services, or personal social services, rather than being separately organized and administered, are often attached to other major social services, such as social security, health care, education, and housing. This is explained by the course of their historical development. The means open to policy-making and administration in the personal social services are often incompatible. For example, the demands of the general integration and coordination of care programs can conflict with the provision of services that take due account of the needs of specific client groups. Also to be reconciled are the provision of individual services and the provision for family and neighbourhood needs.

Statutory and voluntary social services have evolved in response to needs that could not be fully met by individuals either alone or in association with others. Among the factors determining the present nature of such services are, first, that the growth in the scale and complexity of industrial societies has added to the obligations of central and local governments. Second, the increasing wealth and productivity of industrial societies has heightened public expectations regarding standards of living and standards of justice, at the same time augmenting the material capacity to meet those expectations. Third, the processes of social and economic change have grown to such proportions that individuals are increasingly ill-equipped to anticipate and cope with the adverse effects of such change. Fourth, it is difficult and sometimes impossible to recognize and provide for the idiosyncratic needs arising from the interaction of social and personal life.

Any family can experience crises that it is powerless to control. The hardships of ill health and unemployment can be compounded by loss of income; divorce and separation can impede the welfare and development of young children; and long-term responsibility for dependent relatives can impair the physical and emotional well-being of those who provide the care.

A very small number of families experience such intractable problems that they require almost continuous help from personal social services. Some of these families present problems of deviant behaviour, including family violence and child abuse, irregular attendance or nonenrollment in school, alcohol and drug abuse, and crime and delinquency. Not all poor families, however, make heavy demands on social welfare services; indeed considerable hardship could be alleviated through more efficient use of existing services.

Over time, social workers have acquired a special responsibility for people whose particular needs fall outside the aegis of other professions and agencies. Apart from the requirements of individuals and families with serious long-term social and emotional problems, personal social services meet a wide spectrum of needs arising from the more routine contingencies of living. Inevitably personal social services are primarily concerned with reacting to a crisis as it occurs, but today much effort is being invested in preventive work and in the enhancement of welfare in the wider community. In this respect comparison can be made with the traditional aim of social security—the reduction of poverty—and the more ambitious objective of income maintenance (see below *Social security: government welfare programs*).

The organization of personal social services in different societies is extremely variable. Ethnicity and urban deprivation have added new dimensions to need that cut across the traditional client categories of families, children, youth, the sick and handicapped, the unemployed, the aged, and the delinquent. Nevertheless, there are continuities and consistencies in the pattern of needs that characterize these major client groups.

MAJOR AREAS OF CONCERN

Family welfare. Social philosophers and caseworkers generally regard family life as the ideal context for the promotion of social welfare. Family welfare programs seek to preserve and strengthen the family unit through both economic assistance, where available, and personal assistance with a variety of services. Personal assistance services include marriage counseling in most developed countries and in urban centres of developing countries; maternal, prenatal, and infant care programs; family planning services; family-life education, which promotes both the enrichment of family relationships and the improvement of home economics; "home-help" or "homemaker" services providing household assistance to families burdened with chronic illness, handicaps, or other dependencies; and care of the aged through such programs as in-home meal services, transportation, regular visitation, and reduced-cost medicines.

Family services

Child welfare. A paramount concern in all family welfare programs is the welfare of children. Whenever possible, children's services are rendered within the setting of home life. Income assistance to parents may help ensure the basic security of the family structure. Maternal, pre-

natal, and child health-care programs are important in all societies but especially so in those affected by widespread disease and malnutrition; infant and maternal mortality rates are in fact the most basic indexes of child welfare.

Day-care services

The increasing number of working mothers worldwide has given rise to day-care services ranging from simple custodial supervision to educational and health-care programs. In some countries, industries are required to provide such facilities for their employees, in recognition of the changing economic pressures on family life.

While the family unit is imbued with great value by most child-welfare programs, these programs must also address the special needs of unwed mothers and their children, broken families, and children whose families, although intact, are sources of abuse and neglect rather than love and nurture. Attitudes vary greatly among the world's societies toward pregnancy out of wedlock. Historically, social and even physical persecution have been common in some communities, but most modern societies recognize a responsibility toward the welfare of unmarried mothers and their children. In industrial countries, and in some developing countries through private charity, services typically include medical care and delivery and counseling regarding the decision to keep the baby or to give it up for adoption. In many countries institutional homes provide for the care both of unwed expectant mothers and of mothers and babies after delivery, in a setting sheltered from the often rigid strictures of family and community. Procedures of adoption vary considerably worldwide, but arrangements are frequently carried out by social service agencies in cooperation with legal authorities.

Whereas orphans once made up the majority of children living in institutional homes, the number of children who lose both parents through death has been greatly reduced by medical advances. Institutional and foster care are now provided mainly to children whose home lives have been disrupted, permanently or temporarily, by marital discord, financial hardship, parental irresponsibility, neglect, or abuse. While foster care might be considered preferable because it offers the intimate atmosphere of family living, some children, such as those severely affected by parental abuse or emotional disturbance, may adjust more comfortably to the more impersonal environment of an institution. Although it cannot be determined conclusively whether the increasing incidence of reported child abuse is attributable to falling standards of parental care or to improved detection and reporting, much effort has been invested in supervision, social education, and cooperation between personal social services and health care, education, police, and housing authorities.

Youth welfare. The underlying aim of most social welfare services for young people, apart from those services that address immediate basic needs, is to prepare them for the assumption of responsible roles in the adult world. The majority of programs provide adult-supervised leisure-time group activities, which may range from cultural and social events to athletics to hiking and camping. Participation in such programs is high in most European countries. The former Soviet youth organizations, called Pioneers and Komsomol, were the largest in the world. Some programs, such as Boy Scouts, Girl Scouts or Girl Guides, Young Men's Christian Associations, and Young Women's Christian Associations, have spread nearly worldwide, stimulating the formation of similar groups tailored to local needs. In addition to group activity, youth welfare programs also provide counseling and guidance services on a more individual basis to help meet the personal, social, educational, and vocational needs of young people.

Delinquency

While the above services are intended to provide constructive outlets for the energies of young people, there remain many destructive influences in society. Social services have directed increasing attention to the problem of delinquency in an effort to provide alternatives to the traditional juvenile court/institutional methods of control. In some urban areas so-called street workers approach the problem at its source. Recognition of the importance of peer groups in youth behaviour has led to the use of group therapy in many correctional institutions and in communities as a preventive service or as an adjunct to parole.

Welfare of the elderly. The elderly now constitute the largest single client group using personal social services worldwide. In all advanced industrial societies the proportion of infirm elderly is on the increase, and, although they constitute only a small minority of the retired population, their claim on social services is disproportionately heavy. Because social care for the elderly is often labour-intensive, most countries give full support to the promotion of family care and the expansion and rationalization of informal care on a voluntary or quasi-voluntary basis. Services include transportation, friendly visiting, home delivery of hot meals, nurse visitation, and reduced-cost medical supplies. Senior centres sponsor group activities such as crafts, entertainment, outings, and meals on a regular basis. Nursing homes, variously funded, provide medical and custodial care for those who are unable to live independently. Paradoxically, the majority of elderly people lead independent lives, seldom utilizing personal social services. Indeed, fit elderly people are increasingly in demand as a source of voluntary service.

Group welfare. The settlement movement arose in response to the collective needs of deprived urban communities. Settlement houses today, and similar community centres and other organizations, seek to promote the common welfare of local groups that may differ in language, national origin, race, or religion. Whereas, in the United States, attempts were formerly made to Americanize such groups by supplanting foreign traits of language and custom with American ones, the emphasis of educational and training programs has changed; language and other assimilating skills are taught, but the preservation of cultural diversity is also promoted. In addition to educational and cultural programs, settlements may offer legal advocacy, recreational activities, and health clinics.

Refugees and migrants

Throughout the 20th century the resettlement of massive numbers of refugees forced from their homes has placed great demand on social welfare services. In Europe and North America various church denominations have taken an active role in relief and other welfare work for such groups as well as for migrant and transient elements within the general population.

Welfare of the sick and disabled. Serious illness and disability account for many of the problems addressed by social services. In addition to the need for adequate primary care, the ill and disabled also frequently face disruption or loss of income, inability to meet family responsibilities, the long-term process of recovery or adjustment to handicaps, and ongoing care in the form of medication, therapy, and the observance of dietary or other precautions.

In some countries, medical social workers are local-authority social workers who have been attached to hospitals, local general-practice health centres, and child guidance agencies. They provide the counseling and other supportive services required by the physically ill and the disabled and their families. Especially in countries where free medical care is not available to the poor, the responsibility for means-testing gives the workers an additional, advisory role with respect to their clients' financial problems. Personal social services make arrangements for domiciliary care in the form of regular visits from home-helpers and occupational therapists; special appliances and home adaptations are supplied either by personal social services or by health services. In the case of severely disabled people personal social services run day-care centres to provide relief for family care providers and small residential homes for the most dependent disabled when they no longer require hospital care.

Welfare of the mentally ill. The social aspects and consequences of mental illness were recognized early in the history of social work. The speciality of psychiatric social work developed initially as an adjunct to hospital care in urban areas. Such services have also been provided under military auspices, particularly in wartime. In developed countries today the psychiatric social worker serves at all levels of patient care; social casework may contribute to diagnosis and the course of treatment; educational and counseling services help other family members cope with the problems of hospitalization, treatment, and aftercare;

close work with housing authorities and employers can facilitate the readjustment of patients into community life by means of foster care, halfway houses, sheltered workshops, and regular employment.

Personal social services have been a major contributor to the development of community care for the mentally ill and the mentally handicapped. In the industrialized world generally, though less so in Russia, policy calls for a reduction in the number of patients hospitalized on a long-term basis; this goal can be achieved only by returning patients to their families or accommodating them in neighbourhood hostels providing adequate support and supervision. The bulk of this responsibility has fallen on local authorities and voluntary agencies, which provide the professional staff and volunteers. Treatment programs are also increasingly designed to prevent hospital admissions and to avoid compulsory admission in all but exceptional cases.

THE WORK OF THE PERSONAL SOCIAL SERVICES

Social work training. In practice the demand for personal social services does not fall into clearly defined categories. Welfare needs often overlap, and the needs of individuals often affect their families or associates. The range of skills required for effective service provision is equally complex. Inevitably, therefore, opinions differ on the training and deployment of social workers.

In the United States, the United Kingdom, Canada, Australia, New Zealand, Japan, and India the bulk of training is provided in the higher-education system, whereas in France, Germany, Norway, and Sweden it is conducted mainly in separate institutions. Most social workers are employed in either statutory or voluntary agencies; outside the United States very few are engaged in private practice. There is much diversity in their training and deployment, but the role of social workers has broadened, making them individually responsible for a wide range of methods and client groups. In some cases specialized social workers are deployed in teams. Opinions differ on the relative effectiveness of the alternative methods of intervention—direct casework, or counseling, on the one hand and indirect social-care planning on the other. Voluntary and private agencies tend to perform more specialized roles, centred on particular client groups and age groups requiring special methods of care and service delivery.

Administration of services. *Basic organization.* There are marked national variations in the organization and funding of personal social services. To begin with, there are differences in the relative importance of the statutory, voluntary, and private sectors. Second, even if governments are the major contributors, the proportional allocation of funds for the statutory and nonstatutory sectors varies from country to country. Third, there are variations in the relative importance of central, regional, and local governments with respect to statutory funding, policy-making, and service delivery. Fourth, there are also variations in the degree of administrative autonomy granted to the personal social services.

The social service staff

The paid staff of statutory personal social services includes social workers, community workers, social care assistants, home-helps (homemakers), workers who supply mobile meals, occupational therapists, and psychologists working in a variety of field, day-care, and residential settings. Although social workers account for a small proportion of the social service workforce, they constitute the majority of its professional staff. Their job is to provide casework, or counseling, services in cooperation with individuals and families and to engage in tasks of social-care planning, such as seeing to the delivery of direct services in kind and fostering the involvement and support of informal care providers and volunteers. In most industrial societies social workers have more or less exclusive responsibility for mandatory duties related to fostering, adoption, and other work affecting parental rights as well as for the management of substitute home care or residential care for the main client groups. Probation officers act as social workers with a special attachment to the courts, the administration of probation usually being separate from that of other statutory personal social services.

The increasing orientation toward community care calls for social policies that strengthen the association between formal personal social services and informal networks of social care without losing sight of their differences. The formal public, or statutory, sector and the voluntary and private sectors all have paid career staffs whose objectives and management are bound by explicit rules. The primary tasks of the public sector are laid down by statute; most voluntary and private organizations are registered, respectively, as charities and companies. In countries such as the United States, the United Kingdom, Germany, The Netherlands, and Japan formal voluntary and private agencies receive direct or indirect grants from the statutory sector in return for agreed amounts of contracted work. In the developing countries many welfare agencies are internationally organized and jointly financed by charitable donations and government grants.

Informal care

Informal care is spontaneously provided in the context of families, neighbourhoods, and other loosely structured community-based associations. Without these supporting networks the personal social services would be overwhelmed by demand. Consequently they often make small grants to informal self-help groups and supplement the unpaid services provided to dependents by their relatives and friends. Professional social workers and community workers are increasingly deployed in the recruitment, training, and general assistance of informal care providers. Payment for fostering is a long-established practice in many countries, and this policy has spread to the care of other groups such as the handicapped and the infirm elderly.

Personal social services are prime movers in the humanitarian trend toward caring for dependent people in their own communities, to which the high cost of residential care adds an economic incentive. It is evident that there is no clear boundary between the formal and informal sectors of social welfare. Nevertheless, informal care cannot take the place of formal services, the two sectors being mutually supportive rather than alternative sources of social welfare. Formal social services are a matter of legal obligation; their providers and users are normally strangers to each other, whereas informal care is given and received on the basis of personal relationships. Formal services have a wide membership and are delivered on a continuous basis, without regard to personal considerations. Informal care is highly localized and—although it may reflect intense loyalty and devotion—is often less reliable than formal care in the long run because family and neighbourhood networks are vulnerable to personal crisis and social change. Such care also does not usually extend to those without living relatives or other close associates. There are, of course, changes in priority within formal social services in response to trends such as the increasing incidence of reported child abuse, especially in the United States and the United Kingdom, the growing proportions of unemployed and infirm elderly, and the heightened awareness of racial inequality and injustice.

The United States. In the United States the main social assistance and personal social service programs are county- and state-administered, with substantial federal government support. Many programs are delegated to local governments, and voluntary organizations are heavily subsidized by public bodies via contracts for provision of services. The Department of Health and Human Services is the chief federal agency, and each state has a counterpart of this agency. In addition there is a small but popular and growing private market for fee-charging social services that overlaps the voluntary sector.

Federal policies for the personal social services have changed significantly since the 1960s. The Social Security amendments of 1962 put a premium on the role of rehabilitative casework, although states could also include homemaking and foster care. Between 1967 and 1977, however, income maintenance services (Aid to Families with Dependent Children excepted) were regrouped under the Social Security Administration, and primary responsibility for personal social services was transferred to the Office of Human Development. The 1974 amendments to the Social Security Act (Title XX) considerably extended the scope of eligibility for social services, giving priority to

preventive work and positive efforts to improve the quality of life rather than to the traditional focus on poverty abatement. Casework, or counseling, however, lost ground to community-oriented service programs such as day-care provision, mental health centres, and nutrition programs. Problems of child abuse and alcohol and drug dependence have steadily assumed greater importance.

There has been significant growth in employer-sponsored welfare programs in the private sector and service-purchase schemes linking public, voluntary, and private agencies, accompanied by increasing use of paid volunteers. The promotion of for-profit entrepreneurial services and decentralization of funding and policy management from federal to state agencies is intended to diversify still further the mixed economy of welfare that typifies the personal social services of the United States.

Treatment of child abuse

In both the United States and Canada special treatment programs have been developed for the prevention and treatment of child abuse, but lower priority has been given to preschool and family support programs designed to encourage better parenting and child development. The U.S. Child Abuse Coordinating Program set up in 1972 is based on an interservice approach involving municipal and quasi-public bodies, one of which provides the agency officers. American child protection law is extremely complex because of its dual federal and state components, and, although the best interests of children are generally paramount, it is thought difficult to consider them in isolation from those of the parents.

The mental health care legislation of 1970 and 1972 stepped up the funding of community mental health centres in poor areas, but it was not until the Mental Health Systems Act of 1980 that priority federal funding began to reach those with the worst economic or ethnic disadvantages among the chronically ill, the retarded, and the elderly. There is a growing problem of homelessness among the more mobile patients discharged from mental hospitals, who need higher incomes and more social support if they are to resume independent lives.

Social services for elderly American citizens constitute a typical mixed economy of welfare. Amendments to the Older Americans Act of 1965 have led to the establishment of a network of more than 600 Area Agencies on Aging, which are area-wide planning and coordinating agencies. Locally sponsored senior citizen centres provide group meals and counseling, homemaker, information, referral, transportation, educational, legal, and recreational services. There are also a strong volunteer sector and a rapidly expanding private market. Provisions for the frail elderly under Medicaid and Medicare do not include long-term social care, and the poorest groups are dependent on social insurance and social assistance for the requisite finance. Many not-for-profit and for-profit agencies have developed nursing-home and other special housing schemes that are linked to various reverse-equity mortgage options. Nearly three-quarters of all the states have tax policies designed to reduce the cost of independent living for elderly homeowners with low incomes.

The United Kingdom. In the United Kingdom, as a result of the Seebohm reforms of 1970–71, the funding and organization of personal social services are highly centralized at the local authority level. In each local authority a single social services department serves all categories of client and welfare need. In Scotland, however, the probation service is separate. Personal social services are provided from catchment area offices, although some local authorities delegate this responsibility to small "patch" teams serving neighbourhoods. Roughly half of local authority funding comes from the central government; nevertheless, within strict cash limits, the local authorities exercise wide discretionary powers over the organization and deployment of personal social services. Social work training is centrally regulated, and there is only one (general) qualification in professional social work.

Although income maintenance was transferred to the central government in 1948, local-authority social workers continue to provide small cash grants to families with children when shortage of money is deemed likely to cause a family breakdown. In Britain the separation of income maintenance and social work services was part of an overall policy designed to end the historically stigmatizing association between public assistance and social work in particular and the more general association between poor relief and social welfare. It was also hoped that social work and the other personal social services would shed their low status and become more acceptable in all sectors of society. This philosophy was adopted by the Seebohm Report of 1968 and reflected in the Local Government and Social Services Act (1970), but the resources for a truly universal network of services oriented toward preventing problems were not forthcoming.

Status of social work

British child care law developed in piecemeal fashion over a long period. Nevertheless, it places a clear obligation on the local authorities to protect children at risk and to receive them into care when their welfare is at stake because their parents are deemed unable to provide satisfactory care. Under certain circumstances local authorities can assume full parental rights until a child reaches the age of 18. Separate provisions are made for compulsory admission into care through juvenile court proceedings, when children are "in need of care and control" on various defined grounds, or through matrimonial, divorce, separation, wardship, or criminal proceedings. Care orders may also be issued under the Children and Young Persons Act of 1969, as amended by the Criminal Justice Act of 1982, when children or young persons are found guilty of an offense that, if committed by an adult, would be punishable by imprisonment. Observation and assessment centres and secure community homes with educational facilities on the premises are run by the Department of Health and Social Security.

There are strict regulations on boarding out children in care with foster parents, including thorough investigation of prospective homes, frequent inspections, and the keeping of case records. In English law, adoption is an almost complete and irrevocable transfer of a child from one family to another. Adoption orders are made in the Magistrates', County, or High courts, and adoption proceedings can be initiated only by registered, not-for-profit adoption agencies (including local authorities).

Although English law makes extensive provision for the protection of children, personal social services have a well-established tradition of working with children and families on the basis of a cooperative partnership whenever possible. This tradition includes avoidance of recourse to legal intervention or residential care unless it is in the best interests of the children concerned.

British mental health legislation

With regard to the mentally ill and mentally handicapped, the British Mental Health Act of 1959 anticipated the trend toward voluntary treatment and voluntary hospital admission, and legislation in 1982 introduced even stricter criteria for the protection of patients' rights. Since 1983 certain procedures in the admission and discharge of mentally ill patients have belonged to a new category of specially trained social workers. In cases of compulsory detention, patients have a strengthened right of appeal to the Mental Health Review Tribunals, and there are special provisions for the guardianship of certain types of discharged patients. There are still serious deficiencies in community care for the mentally ill or handicapped as well as the elderly and the physically handicapped, but various joint government and local-authority funding schemes have helped to reduce the numbers in institutional care.

Services for the elderly and the physically handicapped account for roughly half of all British local-authority personal social service expenditure, mainly because of the steady increase in the numbers of the frail elderly and the high cost of care for the minority who live in residential homes. Extensive efforts have been made to improve the quality of domiciliary support, but relatives carry the main burden of home care. There are special housing schemes for the elderly sponsored by statutory, voluntary, and private agencies, and a growing number of local authorities employ paid volunteers to visit elderly people and help them with a range of daily tasks. Perhaps the best guarantee of independence in old age, however, is an adequate income from social security (see below *Social security: government welfare programs*).

The formal voluntary sector makes its own important contribution to the care of all the major need groups, although it is heavily dependent on direct and indirect financial support from both central and local governments. Within the voluntary sector the churches have always played a major part in the provision of both community and residential care. Nevertheless as statutory funding has lagged well behind demand, the private market, especially with respect to services for the elderly, has begun to expand.

Australia. In Australia the state governments and the local authorities, with some federal funding, have the main responsibility for personal social services. Each state has a welfare department, usually an amalgamation of the former children's and public relief departments, providing a general range of casework and community services. Some of the municipal authorities also provide welfare services in conjunction with their public health, educational, housing, and legal aid services. In addition there is a well-established tradition of volunteer work that is subsidized by statutory bodies, sometimes provided on a dollar-for-dollar matching basis. Some of the religiously based charities, such as the Brotherhood of Lawrence, the Society of St. Vincent de Paul, and the Salvation Army, are pioneers in work with severely deprived groups.

France. In France personal social services are not administratively autonomous. A variety of social workers and social care workers are employed by other major public services, such as social security, hospitals, community health care, education, housing, and the courts. There are several types of social worker, including the family social worker (*assistante sociale*) and other specialists in child protection, medical social work, and court work; the homemaker (*travailleuse familiale*); child development workers specializing in the care of handicapped children; social allowance guardians with special responsibilities for families in serious financial difficulties; and the community worker (*animateur socioculturel*), who serves neighbourhood groups. Apart from the statutory services there is an extensive network of semipublic agencies (*caisses*) based on trade unions, family associations, and religious denominations, as well as a variety of independent, not-for-profit organizations financed by state grants.

Social work specialists

The French system of child care is explicitly family-oriented. It is based on services financed by the Ministry of Health and the Ministry of Justice, in cooperation with other family income support services. The judicial services are called upon only if parents refuse to cooperate. Social workers are employed in maternal and child health centres and in municipal and family allowance agencies. Special child-protection officers work closely with pediatric nurses in cases of actual or suspected child abuse, and the procedures for removing children from the home and for providing substitute care are in principle similar to those in Britain. Child care services are unified at *département* level, and there is close liaison between the courts and specialized medical services in child protection work.

The reforms of the 1960s and '70s improved the quality of French social services not only for children but also for the mentally and physically handicapped and the elderly. Since the late 1950s domiciliary care and sheltered housing provisions have been strengthened and diversified, objectives that were upheld in the Laroque Report of 1960 and in the provisions of the Sixth (1971–75) and Seventh (1976–80) Plans. The plans specifically referred to the growing need for more trained staff and for more sheltered housing, residential homes, and nursing homes in addition to increased community care and more generous income support within a better-coordinated framework of health and welfare programs at neighbourhood, local, and regional levels. Social care services for the mentally ill are mainly controlled by the health and employment authorities, but the social workers attached to the regional and local *caisses* play a major part in the provision and coordination of community care.

Germany. In the Federal Republic of Germany there is a long tradition of cooperation between the statutory and voluntary sectors and between these formal agencies and the informal networks of family and neighbourhood care. These arrangements exemplify the principle of subsidiarity (the belief that informal care should, whenever possible, take precedence over state intervention) in European Roman Catholic welfare philosophy, although in Germany all the major religious denominations play an important part in social welfare service. The health care provisions of the income maintenance services do not extend to the longer term welfare needs of the elderly mentally ill or handicapped or those of the physically disabled. These are met largely from public aid. About half of the total expenditure on welfare services comes from the Aid for Care program, which channels much of its funding through the larger not-for-profit charitable organizations.

Roman Catholic welfare philosophy

Sweden. The modern Swedish welfare state emerged from poor-law and charitable traditions in which the churches were prominent. Since the years between the two world wars, the scope and funding of statutory agencies have steadily increased. Local authorities, assisted by central government grants, provide most personal social services and a social assistance scheme, in which investigation of needs and means is undertaken by social workers. There has been a trend toward the unification of specialist agencies into local joint welfare boards, but the municipal communes still exercise considerable local discretion in the organization of their services. Although the extensive role of the state in Swedish welfare has elicited much comment, the scale of voluntary effort is equally noteworthy, as it is in Norway and Denmark.

Israel. Israel has a complex system of welfare services distributed by central ministries, with subdivisions for all the major need groups, including services for wounded soldiers and surviving dependents, a Jewish agency with special responsibilities for immigrants, and a universal labour union (Histadruth) with extensive roles in insurance and welfare and a long tradition of mutual aid based on local collectives (kibbutzim) and cooperative villages (moshavim). This has been supplemented by a network of community centres funded by the central and local governments and by membership fees and overseas donations.

Japan. Japanese social welfare provision is uniquely reliant on employer- and work-based social services, although there is also an extensive but relatively underfunded system of statutory local-authority personal social services for the major need groups. Social workers in these municipal agencies are responsible for both discretionary income support and protective social care. In major cities they cooperate with a growing number of voluntary agencies, of which the Minsei-iin is the oldest and largest. As in the case of income support, health care, and housing, access to welfare services for most Japanese workers largely depends on the size and financial stature of the organizations employing them. Although traditional familial ties are still pervasive, they are weaker in the large cities, as a result of social and geographic mobility. At the same time, the number and proportion of the dependent elderly show a marked increase. Accordingly, Japanese policy has turned toward the expansion of statutory services, and much has been done to foster neighbourhood networks of mutual aid that go beyond the traditional notions of kinship and obligation.

Work-based social services

Socialist countries. It is as difficult to make generalizations about social welfare in socialist countries as it is in the case of the democratic societies referred to above. Nevertheless, in the foremost socialist societies the state provides the formal social services, and the workplace and the trade unions play a large part in service management and delivery. In these planned economies, where work is both a civic right and a formal obligation, social assistance for the unemployed is minimal. In the absence of firm data on this area of provision it must be presumed that families shoulder the main financial responsibility for many of the exceptional needs covered by discretionary provision in the West.

There are no professional social workers in China, nor were there any in the former Soviet Union; but social service workers perform similar functions, especially with regard to child protection and delinquency. The erstwhile Soviet Union had a long tradition of nurtured interdependence between the formal social services and a complex

network of mutual aid, lay counseling, and supportive services. The latter were distributed by street, block, and house committees in the towns and cities, by agricultural collectives in the countryside, and by the parallel agencies of the trade unions and the Communist Party.

The Chinese system of social welfare is also strongly based on the industrial or agricultural workplace. Many essential social services, such as health care, are funded from the profits of collective work and administered by neighbourhood committees. Throughout the People's Republic the guiding welfare principles are self-reliance and mutual aid. Although in exceptional cases families receive grants-in-aid to help with care for dependent relatives, Article 13 of the 1950 Marriage Law states that children and parents are jointly responsible for mutual support in hardship and old age. At the same time, extensive and sustained support is given to schemes of mutual support that extend to neighbourhoods and workplaces, and priority is given to the needs of dependent persons without families of their own.

Self-reliance and mutual aid

The trend in the Balkan states has been toward the decentralization of personal social services and the promotion of neighbourhood voluntary work. State-sponsored organizations such as the Alliance of Friends of the Young and the Pensioners' Associations act in conjunction with a growing network of professionally staffed social work centres financed by the 600 communities that are the basic units of local government. Developments similar to these can be seen in the other countries of eastern Europe where, as in China, there is a strong commitment to the expansion of informal provision for family dependents and neighbours.

Developing countries. In former colonies, such as Ghana, Sri Lanka, Jamaica, India, the Philippines, and Francophone Africa, the basic welfare services grew out of modified versions of the European poor laws, charitable and missionary activities, and the introduction of Western juvenile justice procedures. The oldest school of social work in Latin America was founded in Santiago, Chile, in 1925, and the Ratan Tata Foundation established the first Indian school in Bombay in 1936. New training institutions have since proliferated throughout the so-called Third World, many of them sponsored by the United States Agency for International Development.

In developing countries, where formal social services are generally under-resourced, traditional networks of informal care are the main source of assistance in adversity and old age. High rates of migration and unplanned urban growth, however, have weakened these networks in impoverished rural areas and overwhelmed the limited public services in new cities and towns. Indigenous overcrowding and poor housing, unemployment and low wages, and inadequate sanitation and endemic disease are not responsive to Western methods of personal social service intervention. Priority, often within severe economic restraints, must go to major programs of preventive health care, family planning, basic education, income support, and slum clearance. Nevertheless, community development work is also important in these processes of social development. In the poorest rural areas, where the majority of people live at or well below subsistence level, disaster relief is heavily supplemented by international aid agencies such as the United Nations and its associated agencies, including the World Health Organization and the International Labour Organisation (ILO), charities such as Oxfam and the Save the Children Fund, and the governments of richer nations. In the longer term the enhancement of living standards depends on horticultural improvements, reforestation, water conservation, and those irrigation schemes that can be managed within small communities.

Disaster relief

CONCLUSION

It is clear that the processes of economic and social change create new prospects and new hazards for every generation. This requires constant adjustment on the part of the social services. Political considerations and levels of resources largely determine how social services are organized and how responsibility is apportioned between the statutory, voluntary, and private sectors. Even in prosperous societies the scale and diversity of needs is such that the formal social services are obliged to utilize and support informal systems of social care and mutual aid. The idea of the welfare state as a universal provider for largely passive populations has never had any reality in fact, nor much serious support in political theory. There is widespread evidence of a general trend toward the development of closer links between the formal and informal systems of social care, although this might lead to further variation in social welfare services as societies become more sensitive to their indigenous cultural diversity and develop their own responses to change. (R.A.P.)

Social security: government welfare programs

In international usage the term social security has come to mean all collective measures established by legislation to maintain individual or family income or to provide income when some or all sources of income are disrupted or terminated or when exceptionally heavy expenditures have to be incurred (*e.g.,* in bringing up children or paying for health care). Thus social security may provide cash benefits to persons faced with sickness and disability, unemployment, crop failure, loss of the marital partner, maternity, responsibility for the care of young children, or retirement from work. Social security benefits may be provided in cash or kind for medical need, rehabilitation, domestic help during illness at home, legal aid, or funeral expenses. Social security may be provided by court order (*e.g.,* to compensate accident victims), by employers (sometimes using insurance companies), by central or local government departments, or by semipublic or autonomous agencies.

The ILO uses three criteria to define a social security system. First, the objective of the system must be to grant curative or preventive medical care, to maintain income in case of involuntary loss of earnings or of an important part of earnings, or to grant a supplementary income to persons having family responsibilities. Second, the system must have been set up by legislation that attributes specified individual rights to, or that imposes specified obligations on, a public, semipublic, or autonomous body. And third, the system should be administered by a public, semipublic, or autonomous body.

ILO criteria

In its statistics the ILO includes provisions according to which the liability for the compensation of employment injuries is imposed directly on the employer, although such schemes do not strictly meet the third criterion above. For this reason employer liability is included here.

An alternative but wider term for social security in the countries that are members of the European Union is social protection, which includes voluntary schemes not set up under legislation. In some countries the term social security is used in a narrower sense. For example, in the United Kingdom only statutory benefits in cash are regarded as social security. The term social services is used to cover social security; health, education, and housing services; and provisions for social work and social welfare. In the United States the term social security is restricted to the federal social insurance system (OASDI) as distinct from state benefits and "welfare," which in Europe would be called social assistance. In some countries (for example, Denmark and the United Kingdom) the reduction of poverty historically has been a central aim of social security policy, and the concept of maintaining income has been grafted on at a later stage. In other countries, such as France, measures to deal with poverty have been seen as quite separate from the income maintenance aims of social security.

A report published in 1984, prepared by 10 international experts appointed by the director of the ILO, set out the ultimate aims of social security.

> Its fundamental purpose is to give individuals and families the confidence that their level of living and quality of life will not, insofar as is possible, be greatly eroded by any social or economic eventuality. This involves not just meeting needs as they arise but also preventing risks from arising in the first place, and helping individuals and families to make the best possible adjustment when faced with disabilities and disad-

vantages which have not been or could not be prevented. . . . It is the guarantee of security that matters most of all, rather than the particular mechanisms such as contributory or tax financing, the insurance or service model of delivery, or the ownership of facilities (public/private, profit/non-profit) by which that guarantee is given. . . . The means should not be confused with the ends.

Approximately 140 countries have some type of social security scheme. Nearly all of these countries have schemes covering work-related injury and old-age and survivors' pensions. Well over half have provisions for sickness, and nearly half have provisions for family allowances. The least commonly provided schemes are for unemployment, though at least 40 countries have them.

THE RATIONALE FOR SOCIAL SECURITY

Because general social security schemes based on compulsory insurance did not come into being until the last two decades of the 19th century, it has often been argued that social security in its modern form has been a response to industrialization, which caused large numbers of people to become dependent for their security solely on earnings from employment. Indeed many families became dependent on one male earner and thus on his capacity to find work, to undertake it, and to remain in it. Moreover, industrialization led to the migration of people toward centres of work, thus separating them from the support given by the wider family. In addition, the development of compulsory education prolonged the period during which children were dependent on their parents; later the system of enforced retirement created dependency at the other end of life. This situation is contrasted with an often idealized image of the extended rural family with access to land, on which both husband and wife worked, children started work early, and old people continued to work until they became too frail or disabled to do so. On the basis of this oversimplification, some theorists have proposed that social security developed out of a need peculiar to industrial societies and that there is less need or no need for social security programs in the rural areas of developing countries today.

Industrial societies

It is true that support from the extended family, often enforced by local custom and religious beliefs, contributes to the survival of peasant societies. But by no means do all the rural populations of developing countries have access to land, and many people work for wages in agricultural estates and mines. Moreover, peasant farmers are subject to formidable risks of crop failure, quite apart from the risks associated with the shorter average life span that characterizes developing countries. Although there is a need for social security in rural societies, the importance of specific risks may vary from region to region. Moreover, the irregular incomes in cash and kind emanating from agriculture do not lend themselves to the payment of regular social insurance contributions. Thus, what may be lacking in rural societies is the economic and administrative base for providing such security. Furthermore, provision for sickness and old age is not generally seen as the highest priority by peasant farmers overwhelmed by problems of weather and debt.

Agricultural societies

While the advent of industrialization has undoubtedly added to the need for social security by breaking up the extended family and leading to urban poverty, it is by no means the sole reason why the system evolved. Two of the first three countries to make provision for old-age pensions were primarily agricultural societies—Denmark in 1891 and New Zealand in 1898. The Danish scheme was clearly an attempt to alleviate rural rather than urban poverty. And it is notable that the first province in Canada to develop compulsory health insurance (1962) was Saskatchewan, which was overwhelmingly agricultural. These cases indicate that statutory social security may evolve for a variety of reasons. Moreover, it depends to a considerable degree on the economic level attained by the groups that might be covered and the administrative capacity of the country to operate such a scheme. It is certainly the case that, as countries become wealthier, there is greater willingness to defer consumption by paying insurance contributions or taxes.

HISTORICAL EVOLUTION

In many societies charity has been the traditional way in which provision was made for the poor. Charitable giving has been encouraged by many different religions, and in many parts of the world religious agencies have long collected charitable donations and distributed help to those in need.

The imposition of obligations on communities to pay taxes in order to provide for the poor can be traced back for hundreds of years in a number of different societies. For example, part of the function of the Christian tithe or the Islāmic zakat was to provide for the poor. Town poor laws were passed in Germany from 1520 onward, and a law passed in 1530 clearly placed on towns and communities the obligation of sustaining the poor. In 1794 the Prussian states assumed the responsibility of providing food and lodgings for those citizens who were unable to support and fend for themselves. From the 16th century it became recognized in England that there were people who could not find work, and legislation was passed to provide work for the poor and houses of correction for rogues and idlers. From 1598 a clear obligation was placed on parishes to levy local taxes and appoint overseers of the poor in order to give relief to those who could not work and to provide work for those who could. This formed was the essence of the Elizabethan Poor Laws, an early provision of social assistance.

Poor Laws

The Elizabethan Poor Laws were poorly enforced in the 17th century but widely used and liberalized by the end of the 18th century. A new Poor Law enacted in 1834, and reflecting a harsh moral view of poverty, required the poor persons to be admitted to the workhouse so as to receive relief only in kind, with occasional exceptions, but this again was by no means uniformly enforced, though it added greatly to the unpopularity of the Poor Laws. Some U.S. states copied the Elizabethan Poor Laws but exempted recent immigrants. The English Poor Laws were also introduced in Jamaica in 1682 for destitute European immigrants and much later in Mauritius (1902) and Trinidad (1931). In Latin America the Spanish colonists, instead of establishing a public relief agency, gave grants to charities to provide "hospitals" for the poor (*beneficencias*), and the Portuguese promoted lay brotherhoods such as the Misericórdia.

The first general social insurance scheme was introduced in Germany in 1883. The scheme drew upon three types of precedent. The first was the ancient system of guild collection boxes—funds to which each member of a particular trade was required to contribute at regular intervals; such funds were originally used for hospital and funeral expenses and for food and lodging for aged and disabled members. By the middle of the 14th century these arrangements were covered by statutes and regulations. Relief funds were later established by associations of miners. The second precedent was a Prussian ordinance of 1810 that placed on masters a duty to ensure that their servants were given medical attention in case of illness. From 1849 communities could make bylaws requiring both employers and employees to contribute to relief funds, and a law of 1854 introduced compulsory health and accident insurance for miners. The third precedent was the employer's legal liability to pay damages for accidents caused by negligence. As a result of this liability, which was widened in 1871, many employers took out private insurance. The system did not work well because the burden of proof lay with the worker, who normally had to incur high legal costs and delay before he could hope to obtain lump-sum compensation.

Chancellor Otto von Bismarck's 1883 sickness insurance law provided to employees in defined types of industry both medical care and cash benefits during a period of sickness, to be paid for out of contributions from both employees and employers. This was followed by a law of 1884 making accident insurance compulsory. The schemes were operated by numerous funds controlled by the insured and their employers. Finally a law establishing a pension for all workers in trade, industry, and agriculture from the age of 70 was passed in 1889. This was directly administered by the Imperial Insurance Office. Austria

followed part of the German example in 1888, Italy in 1893, and both Sweden and The Netherlands in 1901.

Bismarck's political aim in introducing social insurance had been to address the legitimate grievances of workers so as to check the growth of socialism and avert revolution. A proportion of previous earnings were to be paid in cases of sickness, injury, widowhood, and old age. Employers and employees were to work together in implementing the scheme. In Austria part of the driving force was the Christian Socialists' aim of improving the worker's position. Although Britain had been the first country to industrialize, the developments in Germany and Austria originally attracted little British interest because of an aversion to state intervention, an apparently lesser likelihood of revolution, and the slower development of British socialism.

Friendly societies

In Britain self-help through friendly societies and savings banks was seen as the solution. The friendly societies were run by skilled workers with no employer participation and provided flat-rate cash benefits for sickness as well as treatment by the society's doctor, who was normally paid a flat rate per member insured—a so-called capitation payment. By 1870 membership had grown to 1,250,000 and by the early 20th century to 7,000,000. Apart from the regulation of friendly societies, the only social security legislation passed in the United Kingdom during the 19th century was to widen the liability of employers to compensate workers for personal injury arising out of work. By a law of 1897, compensation could be obtained whether or not the employer had been negligent.

Further action arose in the United Kingdom out of social concern about poverty, which was systematically investigated both in London and in York. In 1899 the government carried out an inquiry into the incomes of 12,000 elderly people. The influential precedents for action were those of New Zealand and Denmark, which had made provision for old age without establishing social insurance schemes, in contrast with Germany, where the scheme was based on insurance. In 1908 in Britain, pensions at age 70 were introduced in a noncontributory, income-tested basis, partly because such a scheme could bring immediate relief to the aged poor, as opposed to a contributory scheme, which could only pay pensions to those who had paid contributions. The social insurance approach was, however, applied to sickness and also to unemployment in certain occupations three years later.

Unemployment insurance

This compulsory scheme, including the first state scheme of unemployment insurance, again reflected Britain's concern to address the main causes of poverty. Benefits and contributions for sickness and unemployment insurance were flat-rate, building on the precedents established by the friendly societies and ensuring the maximum impact on the living standards of low earners. From 1925 the social insurance approach began to be extended to provide for widowhood and old age.

Unemployment insurance was subsequently introduced in Austria and Belgium (1920), Switzerland (1924), Germany (1927), and Sweden (1940). In the case of health insurance, Denmark, Norway, and Sweden promoted voluntary health insurance before making such schemes compulsory, much later than in Britain or Germany. In France voluntary insurance had long been less developed, and mutual insurance societies had long been regarded by government with suspicion, and therefore suppressed. When they ultimately were allowed to expand, around the end of the 19th century, the bulk of their membership was middle class. During the second half of the 19th century larger employers established their own pension and welfare institutions. An employers' liability law was passed in 1898 for accidents at work irrespective of negligence, and in 1910 modest contributory pensions were introduced for industrial and agricultural workers. This law met with limited success, owing to opposition on the part of workers, noncompliance among employers, the loss of rights on change of job or bankruptcy of the employer, and the erosion of the value of pensions during inflation. Health insurance, though provided for in a law of 1920, did not come into effect until 1930, owing to the opposition of the medical profession.

A major innovation came in Belgium (1930) and France (1932) with the introduction of family allowances, although New Zealand had introduced a limited means-tested scheme in 1927. These derived from the ideas of social Christianity regarding "the just wage" and had originally been introduced by Christian employers on a private basis; special funds were later set up to equalize financial burdens among employers. Family allowances became relatively generous in France, partly because of concern to increase the birthrate after the heavy loss of men in World War I. (There is, however, no clear evidence that family allowances have any impact on birthrates.) France later introduced family allowances in many of its colonies during the 1950s.

During the interwar period social insurance schemes were introduced in more and more countries in Europe and Latin America. The most common model was that established in Germany—autonomous funds paying earnings-related benefits. The first group to benefit in Latin America was civil servants, followed by those working in railways and public utilities. There were separate schemes for hospital personnel in Argentina (1921), shipbuilders in Uruguay (1922), merchant seamen in Chile (1925), and dockworkers in Peru (1934). Thus the foundations were laid for the complex social security schemes in Latin-American countries that later reformers tried to amalgamate. The first comprehensive scheme for industrial workers was established in Chile in 1924. In African colonies many schemes of social security were originally introduced only for expatriate Europeans.

The Great Depression of the 1930s finally overcame opposition in the United States to federal intervention in social security. Earlier government activity had consisted of piecemeal initiatives at the local or state level. The Social Security Act of 1935 not only provided federal grants for state public assistance to the aged, blind, disabled, and dependent children but also established a federal old-age insurance scheme and federal financial backing for state unemployment insurance plans that met federal guidelines. Provision for survivors was added four years later and for disability later still. A quite different approach was taken in New Zealand, which introduced in 1938 the first universal non-means-tested pension from age 65, available only on a test of residence and financed in part from a special social security tax on income.

U.S. Social Security Act of 1935

A major influence on world developments was the British government's report by Sir William (later Lord) Beveridge in 1942, which argued for the maintenance of full employment as a responsibility of government, family allowances for all children after the first, comprehensive health care for the whole population, and a unified national scheme of social insurance run by the state with the safety net of a unified national scheme of social assistance. The aim was to eliminate want or poverty. By 1948 the scheme had been introduced in the United Kingdom with some compromises and modifications. A drive, inspired by Pierre Laroque, to unify social insurance in France after World War II was less successful.

During the period of rapid world economic growth from 1945 to 1973 there was a further major expansion of social insurance to more countries, covering higher percentages of population and wider risks. The expansion was particularly notable in Latin America and in certain French colonies in Africa, where comprehensive social insurance schemes were introduced following the original schemes for family allowances. In the British colonies a different approach was taken: provident funds (see below) were widely developed for particular categories of workers. Discrimination on racial grounds was widely prohibited but still persisted in South Africa.

Innovations after World War II

The major innovations in social insurance after World War II were the protection of pensions by linking them to the inflation rate; the development of dynamic pension formulas that indexed past pension contributions to the level of earnings at the time of retirement; the introduction of flexible retirement providing for part pension and part-time earnings in the last few years before full retirement; the movement toward equal rights for men and women; attempts to provide for all disabled people on the basis of the degree rather than the cause of disability (*i.e.*,

whether or not work-related); the growing recognition of extra needs arising from disability and of the needs of persons caring for the disabled; special provisions for one-parent families; the development of parental allowances in addition to family allowances; the integration of child tax allowances with family allowances; and the extension of the same health-care rights to all citizens.

METHODS OF PROVISION

Legal liability. Many countries that once held employers themselves legally responsible for compensating victims of work accidents and for paying for their medical care have now adopted state schemes of compulsory insurance. From the point of view of the worker the problems with the former system include the delays and costs of going to court and the possibility that the employer may be uninsured, unable to pay, or bankrupt by the time the case is heard. Moreover, a lump sum awarded by a court cannot be invested so as to provide a secure inflation-protected income for life. And when the employer is privately insured, the insurance company is in a position to offer the worker a small lump sum soon after the accident, knowing that the worker may well accept it rather than incur the delay, costs, and uncertainty of a court case to obtain the full value of the claim. From the national point of view such a system is wasteful because of the legal costs and the high administrative costs incurred by the insurer and passed along to the insured by way of higher premiums. The argument in favour of this approach is that insurers quote premiums for individual employers according to their experience of risk, which provides financial incentives for industrial safety. But insofar as such incentives are effective, premiums for a national program of accident insurance can also be risk-rated.

Private versus statutory compensation

In some countries, when a statutory insurance scheme of occupational injury has been introduced, the right of the employee to sue the employer for negligence is removed. In other countries the employee is free to supplement industrial injury benefits by making a claim for negligence.

The legal liability approach is still used in many developing countries for the general provision of medical care. Thus large employers or employers of labour in mines or specific agricultural estates (*e.g.*, sugar, tea, and rubber) are required to provide clinics and hospitals for their employees and dependents. This is one way of ensuring that health services are provided to people working far from the main urban health services. It is, however, difficult to ensure that employers comply with the spirit of the law. Moreover, employees may suspect that the doctors and nurses working in such services owe their primary loyalty to the employer and thus tend to economize on the treatment or are reluctant to certify time off for sickness. A further problem is that it is uneconomic to provide government services in these areas for the remainder of the population who are not employed by the major local employer and is difficult to integrate employers' services with government services.

In several countries employers are required to provide defined levels of cash benefits during short periods of sickness (*e.g.*, six to eight weeks). This avoids the administrative complexity of a social insurance benefit paid by a national scheme or a sick fund supplemented by an employer's scheme. Provision may be made to protect the workers' rights if the employer goes out of business.

While social insurance is preferred to the employer liability approach by social security experts because it can give better protection, employer liability is still widely used in developing countries not only for employment injury but also for sickness and maternity benefits and employer severance payments.

Provident schemes. Many developing countries require certain employers to contribute to a provident scheme providing a lump-sum payment in the event of death or disability or on retirement. Such a scheme differs from a social insurance scheme in that each worker usually has his own personal account from which he or she can draw if certain contingencies arise; there is no pooling of risks among members as there is in a social insurance scheme. Such schemes, which avoid the administrative complexity of paying a regular cash benefit, may be a step toward a full-fledged social insurance scheme. There are three disadvantages of such schemes from the point of view of the beneficiary. First, provision is inadequate for risks occurring early in working life. Second, the funds are generally invested in government stock with a rate of interest fixed in money terms that may be below market rates; the real value of the accumulated savings may thus be substantially eroded by inflation by the time of retirement. Third, a lump sum once received cannot normally be securely invested to provide an income protected against inflation. Moreover, it may be frittered away or unwisely invested. From the point of view of governments, however, such schemes are attractive in that they generate forced savings that can be used to finance national development plans.

Lump-sum benefit

Social insurance. The use of compulsory insurance as a mechanism to provide medical benefits and cash benefits in the case of sickness, disability, widowhood, and old age became acceptable to legislative bodies fearful of accepting extended state intervention that would require higher taxes to finance pensions or other benefits. In societies where self-help by voluntary insurance had been widely supported, the further step of compulsory insurance was seen as a means of making workers "good" by legislation. Because the schemes were financed by contributions levied on both employers and employees with, in some cases, modest state subsidies, unacceptable levels of national taxation were avoided; indeed, as such schemes reduced the need for social assistance or poor relief, the burdens on local taxation were reduced.

Reduction of tax burden

Compulsory insurance contributions are essentially a tax on earned income. Employers try—and probably succeed in most circumstances—to shift the burden of their share of the contribution either to consumers in higher prices or more probably, in the long run, to their employees by paying them less in cash. Thus employers' contributions are in most cases not paid at the expense of profits. However, the fact that the worker is told that the employer has to pay a proportion of the total contribution helps to make such schemes acceptable to employees, quite apart from the clearly defined benefits that flow from paying their share. Compared with the complexities of an income tax, a social insurance tax is a simple one to collect. But if the level of contributions is high, it creates incentives for workers to become self-employed in what has come to be called the "black," or "underground," economy and for employers to avoid contribution liability by employing contract labour rather than full-time staff.

In terms of meeting social needs or reducing poverty the social insurance method of provision has a number of disadvantages. Over the years many countries have tried to find means of countering these. First, the analogy with private insurance, which made such schemes politically salable, carries with it the social disadvantage that benefits should be paid to those who have contributed. Thus such schemes cannot provide benefits to persons who have never worked, for example, persons who have become disabled before reaching the age to enter employment, those incurring risks very soon after entering employment, and women (or men) who do not enter the labour force because of family responsibilities. Second, the expectation that benefits should be related to the amount paid in discriminates against individuals, usually women, who because of family responsibilities have fewer years in paid employment. Moreover, workers with dependent spouses and children have greater needs than single persons, though the assumption of marital responsibilities—or the converse assumption of marital dependency—is not strictly speaking an insurable risk. Third, where contributions are related to earnings, the benefit will be low for low earners, thus failing to protect them from poverty. The alternative approach, which some countries have adopted, of flat-rate contributions and flat-rate benefits can impose heavy burdens on low earners with family responsibilities. Fourth, it is difficult to bring the self-employed and those working for small employers (*e.g.*, agriculture or domestic work) into such a scheme.

Over the years many countries that started with a purist insurance approach have modified their schemes to try to

overcome many of these disadvantages. For example, extra benefits have been provided to persons with dependents. Contributions have been credited to persons outside the labour force for reasons of family responsibility, sickness, or disability. Minimum benefits have been introduced above those strictly warranted by low earnings-related contributions, or the benefit formula has been weighted in favour of lower earners. And some countries have made contributions earnings-related or integrated them with income tax while still paying flat-rate benefits.

Benefits to all residents. Because of the disadvantages of the social-insurance approach, some countries have made certain benefits available to all residents and financed them out of taxation. When the benefit is paid on the basis of age it is sometimes called a demogrant. The most common benefit selected for this approach is the family allowance. The underlying philosophy is that provision for children should not depend on whether the parent is or has been in paid employment. Some countries have adopted this approach for pensions or at least for a minimum pension. In some cases this evolved from an earlier provision of an income-tested pension. In other cases this was the only way forward for governments in which the power to levy social insurance contributions did not rest at the federal level. Some countries have more recently applied this approach to provision for the disabled in the form of a minimum benefit based only on the extent of disability. It is increasingly applied to medical benefits on the grounds that all citizens have a right to health care.

Demogrants

Social assistance. The development of social insurance and demogrants has not removed the need for social assistance to fill gaps in provision in advanced societies. Social assistance is based on need and thus requires declarations of income, family size, and other circumstances. Thus it is provided on the basis of a means test that takes into account not only income but also capital; persons with a specific level of savings may be ineligible. Alternatively it may be only income-tested, the income from capital being assessed in the same way as other income. Often those who have been given the task of operating the scheme (*e.g.,* social workers) have been allowed considerable discretion in deciding whether to give assistance and how much to give in certain types of cases. Not all basic rules are known to claimants. The tendency in industrialized countries has been to try to transform assistance into a right with published scales and regulations and opportunities for appeal. With codification has often come standardization and the unfortunate removal of some of the flexibility available under discretionary systems.

Means and income tests

In some countries social assistance plays a residual role, providing a less favourable level of support than is normally available from social insurance benefits. In other countries (*e.g.,* the United Kingdom) social assistance plays a considerable role in supplementing social insurance benefits for those without other sources of income such as sick pay or employers' pension schemes as well as providing for those without rights to benefits (*e.g.,* one-parent families other than widows) or those whose benefits have run out because they are paid only for a specific number of months (*e.g.,* unemployment benefits).

There are disadvantages of the social assistance approach. First, it penalizes saving and earning because income from any source is normally deducted from the assistance that would be payable, and persons with a certain level of savings may be ineligible until they have used them up. Second, it tends to stigmatize the recipient; and third, partly for this reason and partly because of the difficulty of knowing detailed rules of entitlement, there are considerable numbers of people who would be eligible but do not make claims. Partly because of this problem of stigma, social assistance programs are called by a variety of different names in the hope that they will be more acceptable to applicants. For example, the term used is supplementary benefit in the United Kingdom and GAIN (guaranteed income) in British Columbia. Eligibility rules differ considerably from country to country and are usually determined locally rather than centrally. Moreover, schemes are generally financed wholly from taxes—often local taxes. In the United Kingdom, where rules are determined centrally, persons in full-time work are not eligible. In the United States only households headed by a single parent are eligible for the Aid to Families with Dependent Children program, which creates incentives for desertion or fictitious desertion. There are, however, further programs for the blind, the disabled, and the aged.

Criticisms of social insurance

The United States uses what is essentially the social assistance approach for meeting the medical care needs of low-income persons under the Medicaid program. Ireland operates a scheme by which persons with low income can apply for a medical card that gives them more extensive rights to free health care than are available to other income groups. Those with low incomes in South Korea can also apply for cards giving rights to free or nearly free health care.

A number of countries in Europe have developed separate income-tested provisions to help persons with low incomes meet the cost of rent or property taxes. Such housing allowances are available to persons whether in work or not and take account of family composition as well as rent payable.

Negative income tax. Partly because of the stigma attached to social assistance, the difficulty the potential beneficiaries have in understanding eligibility, and their reluctance to apply, it is often proposed that the information provided to the state from income tax returns should be used by the state to determine the need for cash payments to persons with low incomes. The ability to do this depends on persons' being required to make income tax declarations by a certain date however low their incomes, which is not the practice in every country. Canada has a program to supplement on the basis of this information the incomes of persons drawing pensions. This approach is much less appropriate for younger people whose financial circumstances change considerably from year to year and month to month due to sickness, unemployment, job changes, marriage breakdown, remarriage, and so on. People need money when poverty strikes, not after the end of the income tax year.

CASH BENEFIT PROGRAMS

Provisions for cash benefits change from time to time in all countries. Thus no description can be fully up-to-date. The information presented here is chiefly based on the returns made by 140 countries to the Social Security Administration of the United States and published in 1985 as *Social Security Programs Throughout the World.*

Pensions. Three basic types of state pension schemes predominate. The first is a flat-rate pension with no income test. This may be available on a test of residence only or with the stipulation that the person has been employed for some specific period and has paid requisite contributions. This approach is found mainly in Scandinavia and the Commonwealth countries. The second is an income-tested pension. The third, and most common, type is a pension related in some way to earnings during working life. A further complication is that most countries with a flat-rate pension later developed a second tier of pension rights based on earnings during working life. In other words, the first and third principles are combined.

New Zealand pays a flat rate pension; financed from general taxation, to all who meet residence requirements at age 60. The rate for qualified married couples is twice the rate for single people. The rate of pension is quite a high percentage of average earnings. The Netherlands also provides all residents with a substantial pension but at age 65; it is financed from contributions and reduced if contributions due have not been paid in any year. The supplement for a wife of any age is less than half the rate paid to the husband. In Ireland the pension is less generous and only available to employed persons with minimum contributions paid. Australia combines the first and second approaches with a flat-rate pension from age 70 and an income-tested pension from age 60 for women or from age 65 for men.

Several countries in Scandinavia abandoned an early means-tested pension in favour of a flat-rate pension after World War II because of the unpopularity, complexity, and discouragement to savings of the means test. Later

the level of the pension was regarded as inadequate for all except the low-paid, and an earnings-related tier was established on top. Thus flat-rate pensions are provided on tests of residence in Denmark, Finland, Norway, and Sweden. In three of these schemes a married couple receives substantially less than two single persons. In each case the schemes are supplemented by earnings-related pensions. Canada followed a pattern of development similar to those of the Scandinavian countries. The United Kingdom also gradually moved over to a two-tiered pension, but rights to both tiers depend on contributions paid with credits for absence from work for approved reasons; employers' schemes can be used to provide a specified minimum upper tier of pension.

The Scandinavian and Canadian two-tiered approaches have a number of advantages. First, non-means-tested basic pensions can be provided to persons without a contribution record—including the disabled, those who have not worked because of family responsibilities, and divorced or separated wives. There is a similar advantage in New Zealand's scheme. But, in addition, those who have had higher earnings and thus paid higher contributions receive higher pensions with the value of these guaranteed in terms of purchasing power by the government. This reduces the scope for employers' pension schemes in which the purchasing power of the pension finally paid depends on how far the yield of investments has managed to keep pace with inflation.

Contributory pension schemes, when they were first established, were run on much the same basis as private pension schemes. The level of contributions was calculated by an actuary, and a capital fund was built out of which the pensions could be paid. Even if there were no further contributions, the money was intended to be available to pay out pensions to contributors from the accumulated value of their contributions. This arrangement is known as capitalization or fully funded financing. The first scheme in Germany, enacted in 1889 and based on capitalization, covered most employed persons with earnings up to a specified level. The earnings-related contributions were equal for employees and employers, and there was a subsidy from the state to provide the low-paid with somewhat higher pensions than their contributions warranted. A breach with the principles of private insurance was made to allow workers close to retirement when the scheme went into effect to receive higher pensions than their contributions had earned them. This system of "blanketing in" older workers has frequently been used in other countries when new pension schemes have been established.

Capitalization

It was the experience of rapid inflation after World War II that led to a fundamental change in the financial basis of pensions. Instead of the contribution level's being sufficient to build up a large capital fund, it was calculated according to the expected cost of pensions due to be paid over the next few years. This pay-as-you-go method of financing statutory pension schemes, which became the normal arrangement, contrasts sharply with private pension schemes. The latter still have to accumulate capital funds because, unlike state schemes, they have no power to compel future generations to join them. Thus state pension schemes are essentially a "compact between generations." Those at work are compelled to pay to the pensioners of today in expectation, written into the law, that their pensions will be paid by the next generation of workers.

A second major development in pensions began in the late 1950s in response to rapid economic growth. It became recognized that, if pensions were paid out on the basis of the money value of contributions paid in over a working life during which real earnings had been growing rapidly, pensions would amount to a low proportion of earnings at the time of retirement and a still lower proportion of what those at work would be earning 10 or 20 years later. Thus complex formulas were introduced to adjust pensions to the general level of earnings at the time of retirement. West Germany set the pattern in 1957 and was followed by several other European countries—for example, Austria, Switzerland, and the United Kingdom. An alternative approach (*e.g.*, in Italy and some eastern European countries) is to base the pension on the last few years of earnings. As this can be unfavourable to workers whose earnings decline in the later years of working life, some countries (*e.g.*, France) base pensions on the best few years of earnings. The former Soviet Union offered an option of the last earning year or the best 5 consecutive years out of the last 10.

The practice of giving low-paid workers higher pensions than were earned by their contributions and those of their employers, which was built into the original German scheme but later abandoned, has been copied in later schemes (*e.g.*, that of the United States). An alternative or further way of helping low-paid workers is to provide a minimum pension, as in Germany or the United States (though in 1981 the U.S. provision was removed for people not yet retired). This particularly helps women, whose average annual income, despite legal inroads against discrimination, remains well below that of men and whose pension contributions are now likely to be interrupted by leave from work for family responsibilities. A much more common provision is an income-tested social pension as, for example, in Belgium or France.

Another development mainly of the period after World War II has been the automatic adjustment of pensions according to an index of prices or in some cases to the average level of earnings, or whichever is more favourable. Some countries have postponed adjustments or modified their formulas, particularly when prices were increasing faster than earnings.

Pension age

The age at which full pension can normally be drawn varies considerably between countries. In Europe the normal age for men can be as high as 67 and as low as 60 and for women as high as 66 and as low as 55. Some developing countries have still lower pension ages. To some extent pension age tends to reflect the expectation of life in the particular country. The pension age for women, however, is often lower than for men; one reason often cited for this is that husbands tend to be older than their wives, and so the disparity in pension ages permits simultaneous retirement. The arrangement, however, is disadvantageous for women who are retired compulsorily at the lower age after having had less time to accumulate a record of contributions. There is, therefore, a trend to equalize pension ages between the sexes. To do this by lowering the male pension age is expensive; it is for this reason that the European Union has not made this binding on member states in its directive on equal rights to social security.

Some countries have long had provisions allowing the pension to be drawn a few years earlier than the stipulated age of retirement with an actuarially calculated reduction in the pension paid. Such provisions are suited to the more generous earnings-related schemes in which a reduced pension would not normally cause poverty. Ill health is a common reason for early retirement, though many choose this option in order to enjoy retirement while still in good health. There commonly are also provisions by which people who wish to postpone their retirement and continue to contribute can draw a larger pension. In some cases these arrangements have been introduced in the hope of encouraging later retirement, thus modifying the deterioration of the ratio between the employed population and the retired population, which necessitates higher levels of contribution by those still working as the proportion of the pensioned population increases.

Despite the logic of raising the normal pension age in line with an improved expectation of life, changes in schemes of industrial countries in the 1960s tended to lower the age. This trend has continued as the level of unemployment has increased, despite the financial burden this places on the schemes, particularly in the long run. The political objectives of reducing the number of persons recorded as unemployed and creating jobs for younger people have taken priority. Thus a wide variety of complex provisions have been written into pension schemes defining the circumstances in which full pensions can be drawn a few years earlier than otherwise stipulated. This may be allowed to those with many years of insurance (*e.g.*, 35), to those who have been unemployed for a substantial period (*e.g.*, a year), to those who are disabled,

to those with arduous or unhealthy occupations, and to those whose jobs are being released for younger persons. In some countries the pension is income-tested below the normal age. Contrary to this trend for earlier pensions, the United States has raised the future pension age in two steps from 65 to 67 in response to the long-run financial prospects for the pension scheme.

Partial pension

A development pioneered by Norway in 1972, and since followed by more and more countries, was to allow persons aged 67 to 69 to reduce their working hours and receive at the same time a partial pension. This enabled older people to make a gradual transition between work and retirement. The change was made when the pension age was lowered to 67. Sweden followed in 1976 with a provision for those aged 60 to 64. Partial pensions have also been introduced in Spain and, on a much more restricted basis, in the United Kingdom.

In most schemes in industrialized societies there is a limit on what the pensioner can earn without leading to a reduction in pension. Some schemes specifically require the pensioner to leave his or her job on receipt of the pension. These provisions add to the wider pressures leading to the steady fall in the proportion of persons in full-time work above the normal pension age. But the main reason for this trend is the increasing generosity of pensions, both public and private.

Early pension schemes made extra provision for a dependent wife, and more did so between World Wars I and II. This can mean that women's contributions are "wasted" in the sense that the pensions they earn are less than they have a right to as a dependent wife. The greater frequency of divorce and cohabitation has meant that more women are wholly dependent on the pensions they earn in their own right. Moreover, some pension schemes make no provision for a dependent wife (*e.g.*, in Germany, Austria, and Italy). The issue of women's rights to pensions is particularly important in the context of poverty, as women on average live longer than men.

A few countries allow housewives to contribute to pension schemes on a voluntary basis, but few women do so in practice. Others have adopted provisions for the dividing of pension credits between spouses. In the United Kingdom a man or woman can be credited with a full year's pension rights for each year up to a maximum of 20 during the whole of which he or she is caring for a dependent child or disabled relative. These rights are based on the individual's previous record of contributions.

In the early schemes widows over pension age were entitled to a proportion of the pension of their husbands. More and more schemes have been amended to make similar provisions for widows and widowers. Usually survivors can choose between their own personal rights and a proportion of the rights of their deceased spouse, but in Sweden widows can receive both earnings-related benefits on top of one flat-rate pension. This right is available, up to a maximum, to a widower as well as to a widow in the United Kingdom.

Disability and sickness benefits. In most countries provision for occupational injury is the oldest form of social security. The original German law of 1884 provided for workers to receive half pay for four weeks followed by two-thirds pay during temporary disability. In cases of permanent disability two-thirds of earnings from the year preceding the accident were paid out, with a proportion of this pension paid in cases of partial incapacity. Extra provision could be made for persons needing constant attention. The scheme was wholly financed by the employer, who paid insurance contributions, assessed on the degree of risk involved in the employee's occupation, to statutorily established associations. The associations then paid out any benefits.

The British law of 1897 made employers liable for compensation but did not require employers to insure against the risk. Compensation was half the basic pay for up to six months, at which point the claim could be settled by a lump sum. These two very different precedents influenced developments in other countries. Continental European countries tended to follow the German model and the Commonwealth and the United States that of the United Kingdom. An act modeled on the British law of 1897 was passed in India in 1923, though the coverage was small. Moreover, the Belgians, Dutch, and French as well as the British preferred to introduce in their colonies laws imposing liability on employers rather than funded insurance schemes. By the end of World War II most colonies had such laws. These laws were often later augmented or replaced by insurance schemes. Some Scandinavian countries require the employer to insure but allow him to choose his own insurer.

Under an act of 1946 the United Kingdom introduced compulsory insurance through a state scheme with the same rate of premium for all employers. Benefits for incapacity were at a flat rate followed by a disablement pension based on degree of disability to which were added other allowances depending on the situation of the pensioner, including the loss of earnings and need for attendance.

Insurance is now compulsory in most industrialized countries, but the use of private insurance continues in a few countries (*e.g.*, Denmark and Finland) and the majority of U.S. states, while some countries give the employer the right to choose between a public or private insurer. Work-related injuries and an increasing number of occupational diseases lead in nearly all countries to higher benefits and more generous provision than are paid for sickness or injury not arising from work. For example, some countries in western and eastern Europe provide 100 percent of previous earnings as a temporary disability benefit. These benefits normally continue until recovery or the award of a long-term benefit. In most countries loss of earning capacity is a major consideration in the assessment of long-term benefits, and partial disability is more generously treated than in other social insurance programs. The long-term benefits in some countries can also amount to 100 percent of earnings. But the procedure of seeking lump-sum settlements from the courts still remains in some countries and some states of the United States, with all the associated costs, delays, and uncertainties and the difficulty of turning a lump sum into a secure income.

There are three special features of most occupational injury schemes that reflect their historical origins in the employer's liability. First, schemes are normally financed solely by employers' contributions. Second, the right to benefits operates from the very first day of employment. Third, a cash benefit is seen as compensation rather than income maintenance. For this reason dependents' benefits are not normally provided, but there is provision for surviving dependents. In addition, a compensatory benefit may sometimes be paid in addition to earnings or pensions. These features are found only in provisions for sickness or disability that are of occupational origin.

Both employers and employees normally contribute when the scheme is based on social insurance. Some minimum period or amount of contribution is generally required before there is entitlement to benefit. The amount of the benefit may depend on how long contributions have been paid, as for pensions, which is disadvantageous for those disabled early in working life. The main benefit is intended for income maintenance and thus cannot be drawn at the same time as other benefits or pensions with the same purpose. Finally, there is more likely to be provision for a dependent spouse.

Increasing uniformity of disability schemes

There has been a tendency in the period since World War II to bring occupational injury schemes into a closer relation to other social security schemes. Switzerland has always covered work-related and other accidents in the same scheme, established in 1911; New Zealand later adopted the same practice. The separate provisions for occupational injury and other disability raise difficult problems in specifying the distinction. Occupational injury normally has to "arise out of and in the course of employment." Some schemes allow travel to and from work to be considered within the "course of employment," while others do not. There are considerable difficulties in identifying whether certain disabilities (*e.g.*, deafness or arthritis) arise from work, and there are instances in which an injury is only partially attributable to the work situation. Part of the justification for combining the provisions for occu-

pational injury and other disability is to eliminate such ambiguities. In addition there is the social argument that it is wrong to pay different benefits to different people, all of whom have the same degree of disability no matter how or when the respective conditions were caused. The Netherlands is the only country that has responded to this argument. From 1976, unified provision for disability has been made irrespective of cause. Costs of such a program can be substantial if all disability coverage is raised to a level approaching that of the previous, often considerably higher, occupational injury coverage.

In the case of sickness that is not associated with any occupational factors, most industrial countries pay a short-term benefit followed by a long-term pension after periods varying from about six months or less to a year or more. Some countries, such as Austria, Belgium, Germany, and the United Kingdom, place responsibility for paying a benefit on the employer for the early weeks of sickness (though he may be reimbursed), after which the social security fund assumes payment. In some countries benefits may not be payable for an instance of illness lasting less than, for example, three days. In longer periods payment for the first three "waiting" days may be included in the benefit. Doctors' certificates may not be required for short spells of sickness. Benefits may be as high as 100 percent of earnings (*e.g.*, Austria and Belgium) in the early weeks of sickness or subject to a maximum, as in Norway, or for the full 52 weeks, as in Luxembourg. Or the rate may be 90 percent, subject to a maximum (Sweden and Denmark). Other countries normally pay only 50 or 60 percent (*e.g.*, France, Canada, and Greece). The benefit is flat-rate with extra for dependents in Ireland and the United Kingdom (after the eight weeks paid by the employer). The benefit is also paid on this basis in Australia and New Zealand, but it is means-tested. The United States is the exception among highly industrialized societies in that in most states there is no provision for short-term sickness apart from a special scheme for railway employees and social assistance, or welfare. In practice, provision is left for bargaining between employers' and employees' representatives.

Long-term invalidity pensions

Long-term invalidity pensions were included in the original 1889 German pension law (which was the first piece of legislation of this kind) for those who had lost two-thirds of their earning capacity. Many countries followed this model as part of (or as a later development of) their pension laws. In European countries invalidity pensions became payable after full short-term sickness benefit rights had been received. After World War II provisions were made in some countries for those who had considerable partial invalidity. Some countries require persons to have been insured for five or more years in order to be eligible for an invalidity pension, though generally there are means-tested pensions in industrialized countries for those who do not meet these requirements. In Australia and New Zealand those who meet residence requirements are eligible for income-tested pensions.

In countries with earnings-related pension schemes the invalidity pension is often calculated in the same way as the old-age pension. This means that the level depends on the number of years of contribution, though some countries have special concessions to enhance the pensions of those drawing them early in working life. Invalidity pensions may be supplemented by allowances for dependents and for constant attendance and other special needs.

Some countries make special provision for housewives who lack the contribution records that would qualify them for an invalidity pension. One such country is the United Kingdom, though the benefit is low and flat-rate. In Denmark a housewife can receive a substantial income-tested pension in her own right. Another group for which some countries have begun to make special provision is those who have been disabled from birth or before entering the labour market. These groups are provided for in the unified scheme of The Netherlands.

Most countries, however, are far from the position in The Netherlands, where all disabled persons are treated on a similar basis irrespective of the cause of disability. Those whose disability arises out of the work situation are generally most favourably treated; some countries provide full compensation for loss of earnings plus special allowances when required. Those who have paid contributions are generally treated better than those who have not, and often benefits depend on how long contributions have been paid.

Unemployment benefits. While sickness and disability are actuarial risks in that the incidence does not vary greatly from year to year, this is not the case with unemployment. It is partly for this reason that the duration for which unemployment benefits can be paid is limited in most countries or that benefits are reduced after a designated period. A further reason is to induce the unemployed to seek and accept work after benefits end or when they fall, although such work may be less well paid than the individual's earlier work and may provide an income that is lower than the unemployment benefit that has ceased.

Limits of coverage

The payment of contributions plays a critical role in policing eligibility for unemployment benefits; as a result the benefit is not payable to all persons who are involuntarily unemployed. The school dropout who has never had a job or has held one only for a short period is normally ineligible for unemployment benefits. Women seeking to return to work after child-rearing are also ineligible, even though contributions were paid before leaving work. A prospective recipient must normally have held a job from which he has been released immediately before the benefit is claimed, and the individual must establish that he is available for work by registering at an employment office. Normally anyone who has voluntarily left a job or been discharged for misconduct is denied benefits or is penalized.

In some countries the level of unemployment benefits is deliberately set at the same rate as the benefit for short-term sickness (*e.g.*, Canada, Denmark, and The Netherlands) so as to create no incentive for the beneficiary to try to establish eligibility for the higher benefit. In other countries the benefit for unemployment is at a lower level than the benefit for short-term sickness (*e.g.*, Germany, Greece, and Hungary). Some countries that pay an earnings-related benefit for sickness pay a flat rate for unemployment (*e.g.*, Bulgaria and Italy). In Australia and New Zealand unemployment benefits, like sickness benefits, are subject to a test of income. The duration of the benefit varies from 13 weeks in Bulgaria to six months in Hungary, Italy, and The Netherlands and a year in France, Germany, Luxembourg, and the United Kingdom; in Belgium benefits can be continued indefinitely. In many, but not all, countries the unemployed can claim social assistance after their unemployment benefits cease. In several countries in northern Europe unemployment benefit schemes are operated by trade unions, though with substantial government subsidy.

Family, maternity, and parental allowances. While only a few countries had family allowances before World War II and several of the schemes covered employed persons only, with financing by the employer, there was a rapid extension of schemes in the 1940s and '50s. The extension was in large part attributable to the influence of the Beveridge Report in the United Kingdom. Following the British example most of the new schemes in Europe, Canada, and Australasia included all resident children. A second influence was France, which introduced flat-rate family allowances for all children of employed persons in its African colonies—a system also introduced in some Latin-American countries (*e.g.*, Bolivia, Brazil, and Chile). The majority of schemes cover only employed persons, but a minority, particularly to be found in industrialized countries, pay allowances to all residents. The United States is exceptional among the latter countries in making no provision at all except in aid to dependent children paid on a means-tested basis.

Some systems of family allowances are intended to reduce poverty in large families or, particularly in eastern European countries, to increase the birth rate; the rate paid per child increases with the number of dependent children, reaching a maximum rate with the fifth or sixth child and subsequent children in, for example, Australia, Belgium, France, Ireland, and Norway, or the eighth child,

as in The Netherlands. In the former Soviet Union, family allowances began with the fourth child and reached the maximum rate at the 11th. Some systems seem to suggest a desired maximum family size insofar as the rate falls for subsequent children once there are three (*e.g.*, Bulgaria, the Czech Republic, and Slovakia) or two (*e.g.*, Greece and Hungary), or allowances may be payable for a maximum of six children (*e.g.*, Morocco). Finland recognizes that a mother is less likely to go to work if a child is under three years of age and therefore pays a supplement. On the other hand, Austria pays higher rates for older children because they are more expensive to maintain. Entitlement to family allowances ceases when a child reaches a particular age—in most cases the age when compulsory education ceases, though allowances may be continued when a child continues in full-time education or is disabled.

During the 1970s a number of countries decided to abolish income tax allowances for children and make a corresponding increase in the level of their family allowances. It was recognized that the largest beneficiaries from the tax allowances were high-income families with high marginal tax rates, and it was decided that this indirect benefit for children should be fairly shared among all families so as to increase the efficacy of family allowances in reducing poverty. Changes of this kind were made in Australia, Canada, Denmark, West Germany, Israel, New Zealand, and the United Kingdom. Denmark has gone one stage further and removed family allowances from the higher income groups by means of an income test. The United Kingdom has an additional income-tested allowance called the family income supplement which gives further help to low-income families.

It is the general practice for schemes that provide sickness benefits also to provide a maternity allowance starting before the birth of a child and extending for a number of weeks afterward. In some cases the rate of benefit is the same as for a sickness benefit, but in many cases the rate is higher—66 to 100 percent of previous earnings. Sweden has pioneered a parental allowance that can be drawn by the father as well as the mother to encourage fathers to take their turn in staying at home to look after the young child. In some cases a lump sum is also paid on the birth of a child to help pay for nursery goods and clothing.

Parental allowance for fathers

During the 1970s there was a concerted effort in eastern Europe to try to increase the birth rate by increasing the period for which a maternity benefit was paid and by giving credits in the social insurance scheme to mothers who stayed at home to look after a young child. Similar credits are provided in the United Kingdom for persons who stay at home to care for a child or a disabled relative, but the motive in this case is to increase the personal pension rights of those, particularly women, who have accepted family responsibilities.

Benefits for survivors and single parents below pension age. Provision is normally made for a widow below pension age left with a dependent child. Where pensions are earnings-related, the pension for a widow typically amounts to one-half to three-quarters of her husband's pension rights. In some countries the benefit is income-tested or time-limited (*e.g.*, three years in France). Other schemes vary considerably in the extent to which provision is made for widows. Some countries pay benefits providing widows are of a certain age when their husbands die. The age may vary between 40 (The Netherlands) and 55 (France). Some countries only pay the benefit providing the marriage has lasted for a specified period (six months in Greece; two years in France). Other countries pay the benefit to any widow who is disabled or to widows of any age for a short period or indefinitely. Widows' benefits normally cease on remarriage. A widower may be able to claim rights similar to those of a widow if he was dependent on his wife. Some countries extend widows' rights to divorced women. Increasingly, long-term provision for widows without dependent children is being questioned in societies where the trend has been for more and more married women to engage in paid work.

Provision for single parents other than widows is normally left to social assistance where such a scheme exists. Some countries have a special income-tested benefit. In Australia this is at the same level as an old-age pension for a person aged at least 65 but less than 70. In New Zealand it is less than half this rate for a single parent with one child. The problem with either of these arrangements is that a less skilled woman is unlikely to be able to improve her position by taking paid work because earnings lead to a reduction of the benefit or assistance. Denmark pays an extra family allowance higher than the normal rate per child for a single parent. Norway pays an extra allowance as if for one more child. The United Kingdom pays an extra allowance at just over half the level of child benefit.

Variations in provision between countries. All of the industrialized countries have social insurance schemes, and nearly all of them cover the main contingencies discussed above. The United States is exceptional in not providing family allowances, in not providing short-term sickness benefits in the vast majority of states, and in having no general scheme of national health insurance other than for the aged and the poor. The extent of provision in developing countries varies between those that still make provision by employers' liability and those that make provision by social insurance.

BENEFITS IN KIND

Systems of organizing health services or health insurance systems and of paying providers are changed occasionally but less frequently than the detailed provisions for cash benefits.

The first national compulsory health insurance scheme, introduced in Germany under Bismarck's law in 1883, built upon precedents going back many years in the separate German states. Health insurance had developed mainly on an occupational basis and was a requirement for that occupation. The feudal obligation of the employer to his workers was given legislative substance in a society developing national markets, in which the employer without an obligation to pay to a sick fund might undercut the employer who had such an obligation. But the main reason for the scheme, as mentioned earlier, was to try to contain socialist tendencies.

Compulsory health insurance in Germany

The administration of compulsory health insurance was left in the hands of numerous local sick funds operating under legislative regulations. They became jointly controlled by employers and employees and made their own contracts with particular doctors and hospitals for the provision of services. All lower-earning workers were eventually required to be members of a fund. Doctors were paid in a variety of different ways, including salary and capitation. In the course of time there were major protests from doctors excluded from contracts with the funds, and the profession demanded the right for any doctor to undertake health insurance work. The substitution of payment per case and later fee-for-service payment, which the German medical profession fought for and eventually won, was a means of establishing open competition between all doctors wishing to take part in the scheme.

Health insurance was enacted in Austria in 1888 and Hungary in 1891 on a similar basis. A bill to introduce such a scheme in Switzerland was, however, decisively rejected by a plebiscite in 1900. The British Radical politician David Lloyd George visited Germany in 1908 to see the scheme firsthand and subsequently introduced compulsory health insurance for persons with earnings below an upper limit in Britain by a law of 1911. However, the scheme provided only for the services of the general practitioner and the drugs he prescribed; hospital benefits were excluded except for some provision for tuberculosis, partly so as not to disturb the charitable hospitals that provided free care to those in need. Moreover, as a result of pressure from the medical profession, the benefit in kind was administered by statutory committees for each area, which enabled every general practitioner to participate who wished to do so, rather than by the large number of friendly societies that had previously provided medical benefits under voluntary insurance and had made their own contracts with particular doctors. Payment was on a capitation basis, as in the previous friendly society schemes.

The introduction of compulsory health insurance was

considered in Sweden in 1884 and Denmark in 1885, but both countries decided instead to encourage voluntary insurance by government subsidy. Whereas Norway introduced compulsory health insurance in 1911, Denmark did not follow until 1933, and Sweden not until 1955. In these countries public hospitals were well developed and heavily subsidized. A compulsory health insurance law was passed in France in 1920 but, as mentioned earlier, did not come into effect until 1930 owing to disagreement about the local control of the scheme and a dispute with the doctors about the method of payment. Fee-for-service payment was finally substituted for the capitation system originally proposed. Moreover, as the doctors refused to accept the intrusion of any third party between them and their patients, the scheme operated on a reimbursement basis: the patient paid the fee and claimed a refund for the major part of it from the relevant insurance fund. This reimbursement system was adopted later in Sweden, Finland, and Australia (under subsidized voluntary insurance).

Reimbursement system

When health insurance was established in Russia in 1912, insurance doctors were paid salaries and practiced from government-owned premises. This was the pattern adopted in Chile in 1924. Payment of doctors on a part-time salaried basis for work performed on premises owned by the sick fund became the general pattern in Latin-American countries and in Spain, Portugal, and Greece. In most of Europe and Australasia, however, existing hospitals began accepting insured patients when hospital care became covered by insurance; special hospitals for insured persons were built in Spain, some parts of Italy, and a number of countries in Latin America.

The next major step in the evolution of medical benefits was for countries to make them available to the entire resident population, financed wholly by taxation or financed in part by social insurance contributions. This step was taken first by Hungary in 1920 and then by the Soviet Union in 1937. New Zealand implemented similar coverage in a series of steps—free inpatient treatment for the whole population in 1939 and outpatient treatment, free pharmaceuticals, and part payment of general practitioners' bills in 1941, with further steps later on. The United Kingdom established its National Health Service in 1946. Norway made services available to all residents in 1956, Sweden in 1962, Denmark in 1973, Portugal in 1979, and Italy in 1980. By the 1980s more than 20 countries had adopted this system. This does not necessarily mean that all services are free at the time of use. Nor does it necessarily follow that all services are government-owned. Of the eastern European countries, some (Bulgaria, the Czech Republic, Slovakia, Hungary, and Romania) have adopted this approach; others (such as Poland) have not. About half of all countries retain an element of financing by social security contributions after adopting this approach.

Canada was relatively late in establishing compulsory health insurance. The first province to do so was Saskatchewan in 1962. By 1971 all provinces had done so, spurred on originally by a 50 percent grant from federal funds; the provincial schemes became available to all residents. The not-for-profit general hospitals were given budgets by the provinces to provide this care. Australia was also late in changing from subsidized voluntary insurance to compulsory health insurance. The United States and Switzerland are left as the only highly industrialized countries without general compulsory health insurance or a health service available to all residents. There have been many attempts, against strong opposition from the American Medical Association, to introduce compulsory insurance in the United States. However, a limited scheme of compulsory health insurance for the aged (Medicare) was finally introduced in 1966 along with a system of means-tested medical care operated by each state for the indigent and medically indigent (Medicaid).

Some countries in Europe have succeeded in securing high coverage of the population under compulsory health insurance without switching to a service available to all residents. Schemes cover the employed, the self-employed, and all social security beneficiaries and their spouses and dependent children: this can amount to 99 percent of the population. Other European countries (Germany and The Netherlands), nearly all of which have some system of private insurance, exclude the higher income groups from statutory health insurance. Alternatively, some benefits (*e.g.*, hospital care) are available to the whole population, while higher income groups must make their own arrangements for certain other benefits (Ireland).

Apart from Cuba, which has a national health service, only three countries in Latin America (Argentina, Brazil, and Costa Rica) have managed to cover 80 percent or more of the population by health insurance. Moreover, coverage may not necessarily mean that services are equally available. Coverage extends to about half the population in Mexico, Panama, and Uruguay, and more than a quarter in Bolivia and Venezuela. In the remaining countries coverage is 10 percent or less. By no means do all of these countries extend the same rights to the spouse and children of the insured person. Several provide only maternity care and pediatric care for dependents. Coverage is more easily provided for the employees of larger establishments, which tend to be concentrated in urban areas. Even in urban areas those excluded tend to be the self-employed, domestic servants, and itinerant workers. The obstacles to expanding rural coverage include the much lower levels of earnings, the geographic dispersion, the less formal employment conditions, and more extensive self-employment and seasonal employment. Most important of all, some schemes have become too costly to extend on the same basis with tax subsidy to cover the whole population. Thus the remaining population must depend on poorly financed and staffed services provided by ministries of health. Health insurance is, therefore, increasingly criticized for exacerbating inequality in health care by outbidding government health services for trained manpower and for creating a heavy emphasis on sophisticated and expensive curative services in urban areas while the main health need is for preventive services to cut the incidence of infectious diseases in both urban and rural areas.

Japan has managed to avoid the worst of these effects and to achieve high coverage. India, conscious of the damage that health insurance could do to government services, has developed health insurance slowly as resources have become available for doing so in particular states. South Korea has introduced health insurance for the urban employed population and also has provided rights to those with low incomes in urban areas; the problem of covering the remaining half of the population in rural areas remains to be solved.

Many developing countries, particularly those that were previously British colonies, have made health services available to the whole population, providing free or nearly free services. This is the pattern, for example, in the West Indies, Kenya, Zimbabwe, India, Sri Lanka, Malaysia, and many Arab states in the Middle East. In most cases services were originally developed for the expatriate colonialists and extended in the course of time to local residents. The services tend for this reason to be heavily concentrated in urban areas, with little or no coverage of the rural population. With their limited resources, these countries are striving, as part of the World Health Organization's program Health for All, to extend rural coverage with primary health care to all areas by the year 2000.

Among the various national health schemes, benefits are provided in three ways. First is the direct service approach in which the government or insurance fund owns the facilities (hospitals and clinics), pays for supplies, and remunerates the staff on a full- or part-time basis. This is the approach used in the United Kingdom for hospitals and community services and in Scandinavia, where local authorities provide hospitals and clinics, though there may also be a parallel system of doctors working from their own offices. It is also generally used in eastern Europe, in Greece, Spain, and Portugal, in most countries in Latin America, and in most other developing countries. The hospital system in Canada is exceptional; the scheme determines budgets for general hospitals that remain in the hands of not-for-profit agencies.

Methods of provision

The second method is the indirect contract with providers. The providers may be private entities (hospitals or practitioners) or public hospitals, but the health insurance

scheme makes a contract with the provider and pays each provider for services used according to rates established in a negotiated contract. This is the system used for all services in such countries as Belgium, Germany, Luxembourg, and The Netherlands.

The third method is reimbursement, in which the patient pays the bill and applies for reimbursement. The provider may be public or private. This approach is widely used in France, some northern European countries for the parallel system using practitioners in the private sector, and to some extent in Australia and Sweden. The patient may be left to pay part of the bill, as, for example, in France. A fee schedule may be established for rates of reimbursement, but, unless strong measures are taken to prevent it, some practitioners may charge more than the established fee.

In practice many countries use a combination of these systems. Thus, for example, the National Health Service in the United Kingdom, with its direct service provision of hospitals and community services, uses indirect contracts for general practitioners, community pharmacists, opticians, and most dentists. Moreover, where private hospitals are used they are paid under contract, as is also the case in Greece, Italy, and Portugal. In a number of countries in Latin America health insurers use the direct service approach in urban areas but service dispersed populations in rural areas by using indirect contracts.

Health insurance schemes vary in the method by which providers are paid, and this can have a substantial impact on costs. Where doctors and dentists are paid on a fee-for-service basis this provides incentives for the provision of further services—even in France where the patient has to pay a proportion of the cost. In the Common Market countries about twice as many prescriptions are issued to patients when the doctor is paid on a fee-for-service basis as when he is paid on a capitation basis. More surgery is performed where doctors receive fees rather than salaries. Moreover, the patient normally has direct access to specialists and can visit several different doctors in the course of one illness; this also adds to costs. When hospitals are paid on the basis of an itemized bill, more items are often provided. Where hospitals are paid per day of care, there are incentives for the hospital to keep patients for longer than necessary. For this reason, some countries in Europe (Belgium, France, and The Netherlands) have required hospitals paid on this basis to adhere to a predetermined budget. Where hospitals are given a budget from the local or central government, costs are kept under control. Financial incentives for the provision of further services are avoided where doctors are paid on a salary or capitation basis (The Netherlands and the United Kingdom). But this can lead to delays in receiving treatment both for an inpatient and for an outpatient. A provision permitting access to specialists, normally only on the basis of referral by a general practitioner, can be enforced where the patient normally has access to only one practitioner; this helps to limit costs. The system of paying doctors part-time salaries, leaving the doctor free to undertake practice, as in Greece, Portugal, Spain, and most countries in Latin America, can lead to what patients see as poor quality in services—a lack of courtesy and limitation of time devoted to the consultation. For this reason many countries are beginning to offer full-time salaries without rights for the doctor to undertake private practice.

The right to free medical treatment was included in the original German scheme for industrial injury, and provision for rehabilitation was added in 1925. In the course of time more and more emphasis came to be placed on efforts to restore working capacity, and specialized institutions were created for this purpose. Many countries have copied the German example and developed highly specialized institutions owned by sick funds or under the control of the agency responsible for national health insurance for both physical and vocational rehabilitation.

ADMINISTRATION AND FINANCE

Administration of social security. Countries vary considerably in the extent to which their social security apparatus is centralized and unified. A high degree of centralization obtains in the Commonwealth countries and Scandinavia (except for health care and social assistance, which are decentralized to lower levels of government). A centralized scheme may be administered by a ministry or by a semiautonomous agency. In other countries schemes are more often run by separate occupational funds or by funds providing for different risks, as tends to be the pattern in continental Europe and Latin America. The control may rest with boards composed equally of employers and employees. Or it may be tripartite, with the government participating as the third party. In the United States responsibility for social security is divided between federal and state agencies. There have been attempts in some countries to secure greater unification, but such efforts have often encountered strong resistance from particular occupational groups with better benefits or lower contributions attributable to lower risks.

Complexity of regulations

Social security regulations have become extremely complex and difficult to understand. Where there are separate funds, each may have a national office, with no branch offices to which the public has access. Disputes often arise over which fund is responsible for paying benefits to particular claimants. It is, therefore, not necessarily the case that all claimants obtain what they are entitled to receive, and substantial delays can occur while entitlements are sorted out. Problems of this kind are not, however, unique to the public sector. Some private insurance companies are resistant to paying out claims. Unified social security systems with local offices are more accessible to the public, but the offices are not always adequately staffed to give the public prompt and efficient service.

Social assistance regulations are inevitably even more complex to operate than other parts of the social security system. Moreover, they frequently contain a considerable element of discretion. Where schemes are administered by social workers there can be what beneficiaries see as potential coercion; failure to follow the social worker's advice may be thought to lead to the reduction or removal of benefits. Some have argued that all social assistance regulations should be published so that claimants can know their rights and thus be in a position to appeal against decisions to refuse benefits or extra allowances. Adoption of this approach has led in some cases to regulations that are too complex for the staff to operate efficiently, or in others to regulations that have been streamlined at the expense of former provisions for discretion. Particularly contentious is the question of cohabitation. If an unemployed married woman living with her wage-earning husband is not entitled to social assistance, it would seem at first sight only fair that an unemployed woman cohabiting with an employed man should be treated in the same way. But cohabitation may not be accompanied by maintenance and is anyway extremely hard to define. The borderline between a lodger and a cohabitant is by no means clear-cut in all cases nor readily established by any outside agency. Attempts to do so can involve considerable invasion of personal privacy.

The issue of cohabitation

Financing of social security. In most countries the major part of the cost of social security is paid for by proportional contributions of earnings from employers and employees. The contributions may be divided equally between employers and employees, except for the whole cost of the occupational injuries scheme, which falls to the employer. Alternatively the employer may pay about twice the amount falling to the employee. There is usually a "ceiling," or level of earnings, beyond which the contribution becomes flat-rate at the level of contribution due on this maximum of earnings, though this is not the case in either Sweden or Switzerland. The maximum varies from around 50 percent above average earnings (*e.g.*, France, Ireland, and Italy) to twice average earnings (*e.g.*, Germany, the United Kingdom, and the United States) or higher (Norway). The reason for this may be to prevent insurance contributions from overlapping with high marginal rates of income tax or to leave the replacement of high earnings to the private sector. Some countries also exempt very low earners from contributions or make the employer pay them instead of the employee.

Usually some portion of costs is left to be met from taxation. At the very least the government will stand by

to meet any deficit between benefits and contribution income. During the 1970s there was a trend in most countries in western Europe for costs to be shifted away from employers and onto taxes (*e.g.*, Denmark, Ireland, Italy, The Netherlands, Portugal, and the United Kingdom) or to employees (Austria, France, and Germany). One reason for the trend toward tax financing was the growth of unemployment financed by social assistance payments.

Countries in which no costs at all fall on taxes include the small schemes in Burundi and Ethiopia and the wider schemes in Malaysia, the Philippines, and Singapore. At the other extreme, however, countries where contributions play a very small role and by far the bulk of costs is covered by taxation are Australia, Denmark, and New Zealand. In the United Kingdom, where the national health service is primarily financed from taxes and social assistance plays a major role, roughly half of the costs are borne by taxes and half by contributions. Several eastern European countries have no employee contributions; instead, their schemes are mainly financed by employers.

Contributory versus tax-financed schemes

The relative merits of financing by contributions or taxes have long been debated. In favour of contributions it is argued that making beneficiaries pay prevents irresponsible increases in benefits and, where there are separate funds, encourages participation by both employees and employers. The payment of contributions also helps to ensure that commitments are honoured. Contributions are administratively easy to collect since the employee has an interest in securing compliance by the employer. The benefits to the employee of paying are clearly identified, while the cost falling on employers may create some incentive to prevent certain occupational risks from arising. Finally, only by earmarking contributions can earnings-related benefits be justified.

The critics of contributions argue that where they are flat-rate or where earnings-related contributions are only payable up to a low ceiling of income they are regressive and constitute a heavy burden on the poor; progressive taxes on income would be preferable, as they vary according to ability to pay and are also levied on investment income. It is also argued that tax-financing enables governments to judge priorities among all fields of public expenditure, and, where it leads to administration by government, this secures closer coordination between social security and other services. In addition, high contributions lead to the growth of the black, or underground, economy. This is a major problem in France and Italy with their high employers' contributions and leads to a widespread lack of social insurance coverage.

An argument that became more strongly pressed when levels of unemployment rose in the 1970s was that high employers' contributions made products uncompetitive in world markets, particularly in the case of labour-intensive industries, compared with products from Third World countries where social security is less developed. This was said to sharpen the recession and aggravate unemployment in highly industrialized countries. While it is true that employers might gain a short-term advantage if contributions were lowered, it is much less certain that this gain would be sustained in the long run. What was gained in lower contributions might sooner or later have to be conceded in higher wages and salaries or in other wage costs. If the argument were valid, such countries as Australia, Denmark, or New Zealand, which make little use of employers' contributions, would be seen to be cornering a heavy share of world trade. The fact that this has not happened reinforces the argument that it is total labour costs, of which social security contributions are only a part, that affect competitiveness.

It has been claimed that high employers' contributions particularly damage labour-intensive firms and encourage the replacement of labour with capital. In examining this assertion it is relevant first to remember that firms making capital goods also have to pay the same high employers' contributions and that capital-intensive firms pay them indirectly on raw materials, facilities and equipment, and energy. Second, high employers' contributions may well cause cash wages to be lower than would otherwise be the case so that total labour costs are not, in fact, increased by employers' contributions. Third, insofar as high employers' contributions encourage all firms to use more capital-intensive methods of production, this applies to labour-intensive firms as well. This encouragement of investment may lead to production at lower cost and thus a more competitive position in world markets in the longer run.

While there is a lack of convincing evidence that employers' contributions are bad for employment, a low ceiling on contributions may itself damage employment. It may discourage offers of part-time work and lead employers to prefer offering overtime to taking on additional workers. This was the view of international experts appointed by the International Labour Office, who therefore recommended in their report of 1984 that contribution ceilings be abolished.

The substitution of taxes for contributions may not relieve poorer workers if the extra taxes come from goods such as tobacco that are consumed more heavily by those with low incomes than those with high incomes in industrialized countries. There is no guarantee that governments would raise the extra revenue from progressive taxes; they may, for example, lower the threshold at which income tax is paid.

The argument in favour of contributions

The strongest case for contributions is that they justify earnings-related benefits. The strongest case for taxes is that they are used in many countries to make benefits available to all residents—whether the benefits be health care, family allowances, or minimum flat-rate pensions. Solutions to the problem of persons not currently covered or inadequately covered by social insurance programs normally require a greater element of tax financing. This has been the trend in many countries.

The rising cost of social security. The cost of social security rose substantially in the period after World War II both in real terms and as a proportion of rising gross domestic product. While social security spending amounted to less than 10 percent of the gross national product in nearly all countries in 1950, it had risen to 20 to 30 percent or more in many European countries by 1980. Among the reasons were the extension of the coverage of social security, the widening of the risks covered, the indexing of benefits, and the greater generosity of benefits, which moved up to or near 100 percent replacement of earnings for certain contingencies in some countries. But also of major importance was the maturing of pension schemes. Many of them were recast in the 1940s and '50s, and therefore it was not until the 1980s that people had had the opportunity to contribute on the new basis for all or most of their working lives and thus could draw pensions approaching or reaching the maximum for which these schemes provided. Three further factors were the increasing proportion of aged persons in the population, the decline in pension ages, and the lower proportion of working population.

The costs of health care also rose sharply after World War II. Several reasons contributed to this trend. First, the higher proportion of elderly in the population influenced health care costs as well as the costs of cash benefits. Persons over pension age require two to three times more health care than persons of working age, and the difference is still greater for those over 75, the fastest growing age group. A second factor was the decline in working hours, which meant that more persons (*e.g.*, nurses) were needed in order to staff 24-hour services. A third factor was the continuous development of medical technology, such as new equipment and labour-intensive procedures. Instead of replacing labour, as in industry, innovations in health care normally required more labour for their operation. A further reason was the removal of supply restraints with the provision of more doctors and dentists, a major growth of medical auxiliaries, and the construction of new hospitals, which were more expensive to run. A fifth reason was the financial incentives to supply more services, which underlay many of the systems of paying providers under health insurance.

The final and critical factor that destabilized the finances of social security schemes was the rapid growth of unemployment beginning in the 1970s. In those countries

that included unemployment benefits in their social insurance schemes, this phenomenon created both unpredicted higher costs for benefit payments and a loss of revenue from those who were unemployed. The burdens on social assistance programs were also substantial in some countries, coming at a time when unemployed persons were no longer in a position to contribute to tax revenue.

The rapid growth of social security expenditure attracted little attention during the period of rapid economic growth up to 1973. It began to cause concern after the steep rise in oil prices checked economic growth in oil-importing countries. The revenue that financed social security ceased to be buoyant at the same time as new major demands were made on the system. From the late 1970s there was talk of a crisis in social security financing.

By 1980 social security expenditure amounted to 32 percent of the gross national product in Sweden, between 25 and 30 percent in Belgium, Denmark, France, and The Netherlands, and between 20 and 25 percent in Austria, West Germany, Ireland, Luxembourg, and Norway. These figures were much higher than for Australia (12 percent), Canada (15 percent), Japan (11 percent), New Zealand (14 percent), the United States (13 percent), or the United Kingdom (18 percent). The cost was much lower in developing countries. High costs are associated with high levels of social security benefits and also with costly systems of providing health care. Some countries, such as Sweden, have allowed health care costs to continue to rise because of the capacity of this service sector of the economy to provide further jobs and thus avoid high rates of unemployment.

Attempts to contain costs

The aim in many industrialized market countries came to be the containment of the costs of social security. This requires that program costs not grow faster than the yield of contributions. Various devices were introduced to help secure this result. Systems of indexing benefits and pensions to prices or earnings were revised downward, or adjustments were made less frequently. Pensioners were made to pay contributions toward health-care benefits. In France tax income was brought in to supplement the yield of contributions. In the United Kingdom the earnings-related additions to short-term benefits were abolished.

A series of measures was introduced to limit the cost of health care. Charges and copayments were increased or new charges were introduced. Payment for drugs was introduced in West Germany (1977), Italy (1975), and Portugal (1982). Portugal and Luxembourg joined France and Belgium in charging for consultations with doctors. Charges for hospital care were introduced or extended in Belgium, West Germany, Portugal, and France. By 1984 there was no country in western Europe that provided free care to all its insured population.

Payment systems under health insurance were revised to reduce incentives for overservicing. The aim in West Germany was to pay the doctor more for the consultation and less for medical procedures. Payments for diagnostic tests were sharply reduced in Belgium. As part of the introduction of a national health service in Italy, payment to all general practitioners was changed from fee-for-service to capitation, and the bulk of specialists began to receive full-time or part-time salaries. Budgets for each hospital were introduced in Belgium, France, and The Netherlands, in part to discourage unnecessary retention of patients paying per day of care. Countries in which hospitals were already paid on a budget basis reduced the budgets. In the United States hospitals began to be paid under Medicare and Medicaid according to a schedule of costs for various groups of diagnoses.

Countries maintained strong controls over new hospital construction or expansions, and incentives were created in a number of countries to transfer beds from general use to the care of the long-term sick. Several countries took measures to develop alternatives to hospital care, such as outpatient surgery, outpatient hospitals, nursing homes, residential homes, and home care by domiciliary teams. The United Kingdom closed some 400 hospitals over a period of 10 years. Restrictions on the installation of major new medical equipment went into effect in Belgium and France. By 1955, 10 of the 12 countries of the European Economic Community had instituted quotas for medical schools. In Denmark, France, Ireland, Portugal, and Spain the number of medical students was cut substantially.

Most countries in western Europe introduced restrictions as to what medications a doctor could prescribe under the health service or health insurance system. Most of these countries exercised tight control over pharmaceutical prices and pharmacists' margins. New measures were introduced in the effort to control overprescribing.

Variations in social security spending

Social security spending tends to vary between countries in direct proportion to their respective standards of living; in other words, the more affluent a country is, the more it is likely to spend on social security. Spending also tends to vary according to the proportion of elderly people in the population. Third, it varies according to the year in which the first legislation was adopted: countries with older social security programs tend to spend more. There are, of course, exceptions to this pattern. For example, the United States and Japan are low spenders both for their standard of living and for their proportion of elderly, and New Zealand is a low spender for a country that introduced pensions as early as the end of the 19th century.

This type of analysis has been criticized, however, for ignoring private arrangements, particularly employers' provisions established as part of collective bargaining. Thus, for example, the large role of fringe benefits in Japan helps to explain the relative lack of development of statutory social security. Similarly, the large role of occupational pensions and health insurance negotiated between employers and employees helps to explain the underdevelopment of statutory social security in the United States. Hence it is argued that private and public social security must both be taken into account in any comparison of national programs. In federal countries such as Australia, Canada, Switzerland, and the United States there were constitutional obstacles to adopting social security that led to the private sector's playing a larger role.

Political orientation also plays a role in explaining the extent to which social security has been developed in the public sector. After some initial opposition, political parties drawing substantial support from the working classes and the trade unions have promoted the expansion of social security. This includes European Catholic Workers' movements. Extensions of social security may be introduced by coalition governments with a conservative majority as the price needed to keep the coalition together. The high spending in Scandinavia can be explained by the strong influence of social democratic parties in the period following World War II. Trade unions have had less influence in this direction in Australia and New Zealand. The absence of a working-class party in the United States is part of the explanation of the relative underdevelopment of its social security program.

Implications

Some of the trends leading to increased costs are bound to continue. While the number of aged persons in most highly industrialized societies is likely to stabilize during the later years of the 20th century and the early years of the 21st century, the proportion within it of those over 75 will continue to increase substantially. This has major implications for further increases in the cost of health care. Moreover, pension schemes are still maturing, and there are pressures for further improvements of benefits, particularly to provide sex equality, lower pension ages, and better assistance for persons, particularly women, inadequately provided for previously. On top of all this, costly developments in medical technology continue. If the trend to shorter working hours continues this will also have a further major impact on the cost of health services.

Looking further ahead, the proportion of aged in the population is expected to start to increase substantially in the second and third decades of the 21st century as the increase in births after World War II becomes reflected in an increase in pensioners. It is this prospect that has led the United States to plan for increases in pension ages and the United Kingdom to decide to scale down its second tier earnings-related pension scheme.

The level of contributions and taxes needed to sustain present plans for social security cannot be predicted. While the continuing trend toward a higher number of aged in

the population can be safely predicted, the birthrate is much harder to forecast. Of vital importance is the level of unemployment because of its impact on both sides of the balance sheet; reduced unemployment would add to contributions and tax income as well as lower the cost of benefits. Nevertheless, the prospect of a substantial increase in pensioners in the 21st century has led to fears in some quarters that the "compact between generations" may not perpetually be honoured. Hence it is argued that the pay-as-you-go method of pension contribution should be replaced by the capitalization method used in early pension schemes and in the private sector. Alternatively it is argued that the privatization of social security pensions would lead to higher savings and investment out of which future pensions could be paid. The disadvantages of either of these approaches are that there would need to be an immediate increase in contributions to provide the planned level of pensions. This could lead to pressure for higher cash earnings. Moreover, the level of pensions would no longer be indexed but would depend on the yield of investments.

CRITICISMS

It has been argued that the high cost of social security is in part responsible for the low levels of economic growth in industrialized societies since 1973. The argument takes three forms. First, it is said that high levels of unemployment benefits reduce the incentives to take paid work. Second, resistance to the payment of taxes and contributions leads to wage demands, inflation, and government deficits. Third, it is argued that because people have rights to social security benefits they are less likely to save; this lowers investment and thus economic growth. For all these reasons social security is said to have contributed to or even to have been responsible for not only low growth but also for high levels of unemployment.

In response to these criticisms it has been pointed out that empirical investigations lend little support to the contention that people prefer benefits to work, though the availability of benefits may make them less willing to take low-paying jobs. Second, it is argued that tax resistance would apply whether pensions were provided in the private or in the public sector. Indeed, if pensions were provided in the private sector they would have to be capitalized, which would require higher contributions and therefore lower cash earnings, leading to still greater pressures for higher pay. Third, the evidence that social security reduces savings is by no means conclusive; indeed, in many countries there has been a boom in the variety of different forms of saving following the establishment of pay-as-you-go systems of financing social security. Moreover, investment is limited by the availability of profitable investment opportunities rather than by any shortage of savings. And if low savings does limit investment, governments can generate a budget surplus out of which investment can be financed.

Some critics argue that social insurance benefits should be replaced by a negative income tax. As countries get richer, it is argued, an increasing proportion of the population is in a position to take out private insurance against the risks for which social security provides. If social security were concentrated exclusively on the lower income groups, provision could be more generous and the burden of public provision could be reduced. The administrative and other problems of using annual tax returns as the basis of making cash payments to individuals whose circumstances are constantly changing as they go in and out of employment, marriage, and cohabitation are considerable. But in the context of saving, it was because income-tested pensions were thought to be damaging to thrift that many social insurance programs were established in the first place. Low earners are unlikely to save if the yield of their savings leads to a dollar-for-dollar reduction in pension. Moreover, many countries in Europe already have income-related housing allowance schemes that serve much the same function even though they are separate from income taxes. Where this is the case, there is no room for a further negative income tax or other income-tested scheme without imposing extremely high tax rates on increased earnings. Most important of all, it is by no means clear that the economically securer members of the population would be willing to accept anything like the existing level of contributions and taxes if they stood to gain nothing from social security. As a result, provision for the poor might be no better or, more probably, it might even be worse than it is as part of a scheme to which all contribute and on which all are in a position to make claims. Historically, services for the poor have always tended to be poor services.

Effects on savings

Empirical studies have shown some small association between higher levels of social security spending and lower rates of economic growth. But it is not clear that one necessarily causes the other. Many other factors are at work. Countries that had a high proportion of the population in agriculture were in a favourable position to achieve high growth in the postwar period as their agricultural populations declined. Moreover, countries that have had relatively low rates of economic growth, such as the United Kingdom and the United States, are relatively low spenders on social security, while countries that have had high rates of economic growth, such as Belgium, Denmark, and The Netherlands, are high social security spenders.

Critics of social security are not confined to those concerned about its effects on the economy or about personal freedom regarding the extent to which and methods by which individuals provide for their security. There are also those concerned about the "target inefficiency" of social security, or its limited redistribution to the poor. This attack is generally directed at earnings-related benefits. Proposals have been made to use the yield of social security contributions, supplemented by taxes, to provide everyone with a minimum income on which they could live at a modest level supplemented by earnings if they wished to take paid work. Social dividend schemes of this kind are seen not only as a way of redistributing income but also of reducing unemployment. A major problem with such schemes is the high level of contributions and taxes needed to finance the minimum income if it is to go to those with jobs as well as to those without them and be sufficient to live on, if only at a minimum level.

Low target efficiency

It is true, however, that high spending on social security has failed to solve the problem of relative poverty in industrialized societies. Yet abolishing poverty was never the original intention, or at least it was not in many countries. Social security was seen as a system of maintaining income by redistribution from the well to the sick, from the young to the old, and from those with jobs to those without them. This involves substantial redistribution but not necessarily redistribution from rich to poor. For instance, while there is more illness and unemployment among low earners, higher earners tend to live longer and thus draw pensions for a longer period.

A study financed by the European Economic Community showed the extent to which poverty remained in the Common Market countries around the year 1975. Poverty was defined as half the average level of living for each country. The study showed that the greatest success in combating poverty was achieved in The Netherlands and, second, in the United Kingdom, though Belgium and West Germany came close to the United Kingdom. Both of the first two countries have primarily flat-rate benefits. Poverty, however, was at its greatest in Ireland and Italy—both countries with substantial agricultural populations—though Ireland also relies on flat-rate benefits.

A variety of reasons explain why poverty persists in industrialized countries despite their elaborate provisions for social security; the precise reasons and their relative importance differ from one country to another. One reason may be that the arbitrarily selected poverty line is just above the level of living provided by social assistance. A second may be the failure of all those entitled to benefits to claim them. A third may be that certain categories of people in some countries are not entitled to claim social assistance (*e.g.*, the long-term unemployed). A fourth reason may be that earnings-related benefits do not secure a sufficient income for low earners to rise above the poverty line or that their record of contributions is insufficient to achieve this result. But considerable poverty may also

persist among families headed by a full-time worker. This is more likely to occur in one-parent families when the earner is a female with limited skills and low earnings. But it may also occur in two-parent families with one earner and several children, where family allowances fall below the cost of maintaining a child at the minimum level or the cost of rent is considerable.

In the case of developing countries social security is criticized for reinforcing the dichotomies between urban and rural populations in general and between employed and unemployed persons in the urban sector. Social insurance contributions, which are in effect taxes specifically tied to providing benefits to the members of the schemes, cannot be used by governments for the benefit of the community as a whole. This limits the capacity of governments to raise tax revenues for broader purposes. Moreover, different health benefits as well as cash benefits may be provided for different occupational groups, thus perpetuating and accentuating inequalities. Those who defend social security argue in reply that requiring part of earnings to be put into an insurance fund is robbing no one outside the scheme, since only by providing the benefits could the particular taxes be justified and compliance with paying them be secured. Criticisms regarding inequality should be directed at the pattern of original earnings, not at social security, which mobilizes part of them for good purposes.

(B.A.-S.)

BIBLIOGRAPHY

Social welfare: Key texts from the extensive British literature include the *Report of the Committee on Local Authority and Allied Personal Social Services* (1968), known as the Seebohm Report; ERIC SAINSBURY, *The Personal Social Services* (1977); *Social Service Teams: The Practitioner's View* (1978); and EILEEN YOUNGHUSBAND, *Social Work in Britain, 1950–1975: A Follow-Up Study,* 2 vol. (1978). The recent historical development of these services is analyzed in JOAN COOPER, *The Creation of the British Personal Social Services, 1962–1974* (1983); and *Social Workers: Their Role & Tasks* (1982), known as the Barclay Report, which brings together different views of the role of the personal social services. *The Future of Voluntary Organisations* (1978), known as the Wolfenden Report, gives a clear account of voluntarism in Britain; and HUGH W. MELLOR, *The Role of Voluntary Organisations* (1985), provides more up-to-date analysis. MARTIN BULMER (ed.), *Neighbours: The Work of Philip Abrams* (1986), studies the role of informal care and reviews the debate on the relationship between the formal and the informal sectors. See also RUDOLF KLEIN and MICHAEL O'HIGGINS (eds.), *The Future of Welfare* (1985).

There are a number of useful comparative studies, including BARBARA N. RODGERS, ABRAHAM DORON, and MICHAEL JONES, *The Study of Social Policy: A Comparative Approach* (1979), on the United Kingdom, France, Israel, and Australia; ALFRED J. KAHN and SHEILA B. KAMERMAN, *Social Services in International Perspective: The Emergence of the Sixth System* (1976, reissued 1980), which covers Canada, the United Kingdom, France, West Germany, Poland, Yugoslavia, Israel, and the United States; SHEILA B. KAMERMAN and ALFRED J. KAHN (eds.), *Family Policy: Government and Families in Fourteen Countries* (1978); JOAN HIGGINS, *States of Welfare: A Comparative Analysis of Social Policy* (1981); NEIL GILBERT, *Capitalism and the Welfare State* (1983, reprinted 1985); J.A. YODER (ed.), *Support Networks in a Caring Community: Research and Policy, Fact and Fiction* (1985); and ELSE ØYEN (ed.), *Comparing Welfare States and Their Futures* (1986). See also CATHERINE JONES, *Patterns of Social Policy: An Introduction to Comparative Analysis* (1985).

Comparative studies on socialist policies include VIC GEORGE and NICK MANNING, *Socialism, Social Welfare, and the Soviet Union* (1980); JOHN DIXON, *The Chinese Welfare System, 1949–1979* (1981); and BOB DEACON, *Social Policy and Socialism: The Struggle for Socialist Relations of Welfare* (1983).

Key issues on personal social services in developing countries are discussed in JAMES MIDGLEY, *Professional Imperialism: Social Work in the Third World* (1981); and MARGARET HARDIMAN and JAMES MIDGLEY, *The Social Dimensions of Development: Social Policy and Planning in the Third World* (1982).

Several of the following studies of social security, notably those of Wilensky *et al.*, Köhler and Zacher, Jones, Easton, and Midgley, include material on the personal social services.

Social security: Nearly all countries make periodic reports summarizing their social security provisions; these are published every few years by the UNITED STATES. SOCIAL SECURITY ADMINISTRATION. OFFICE OF RESEARCH, STATISTICS, AND INTERNATIONAL POLICY, with the title *Social Security Programs Throughout the World.* A second basic international source is *The Cost of Social Security,* published irregularly by the INTERNATIONAL LABOUR OFFICE, which contains comparative tables of the costs, benefits, and financing of social security schemes. For the Common Market countries provisions are periodically summarized in *Comparative Tables of the Social Security Systems in the Member States of the European Communities: General Systems,* published by the COMMISSION OF THE EUROPEAN COMMUNITIES.

A summary of the findings of comparative research is found in HAROLD L. WILENSKY *et al., Comparative Social Policy: Theories, Methods, Findings* (1985). For economic aspects see L.D. MCCLEMENTS, *The Economics of Social Security* (1978). The International Labour Office report by 10 experts referred to in the text is *Into the Twenty-First Century: The Development of Social Security* (1984).

The historical evolution of the schemes in Austria, France, Germany, Great Britain, and Switzerland is described against a wide economic, social, and political background in PETER A. KÖHLER and HANS F. ZACHER (eds.), *The Evolution of Social Insurance 1881–1981* (1982). For a penetrating comparison between Britain and Sweden see HUGH HECLO, *Modern Social Politics in Britain and Sweden: From Relief to Income Maintenance* (1974). For the United States see BRUNO STEIN, *Social Security and Pensions in Transition: Understanding the American Retirement System* (1980, reprinted 1983). Other historical studies of advanced societies include M.A. JONES, *The Australian Welfare State: Growth, Crisis and Change,* new ed. (1983); DENNIS GUEST, *The Emergence of Social Security in Canada,* 2nd ed. rev. (1985); and BRIAN EASTON, *Social Policy and the Welfare State in New Zealand* (1980). For the evolution and current role of social security in developing countries, see JAMES MIDGLEY, *Social Security, Inequality, and the Third World* (1984).

Descriptions of health insurance and health services in various western European countries are given in BRIAN ABEL-SMITH and ALAN MAYNARD, *The Organization, Financing, and Cost of Health Care in The European Community* (1978); and BRIAN ABEL-SMITH, *Cost Containment in Health Care: The Experiences of 12 European Countries, 1977–83* (1984). For other countries, see MICHAEL KASER, *Health Care in the Soviet Union and Eastern Europe* (1976); LEE SODERSTROM, *The Canadian Health System* (1978); SIDNEY SAX, *A Strife of Interests: Politics and Policies in Australian Health Services* (1984); and *Medical Care Under Social Security in Developing Countries* (1982), papers from a meeting of the International Social Security Association.

(R.A.P./B.A.-S.)

Modern Socio-Economic Doctrines and Reform Movements

Since the 17th century, and more markedly since the 19th, political life and thought have been affected by systems of belief often characterized as ideological, by which is meant that they derive programs of practical action in the political sphere (understood to include broadly socio-economic concerns) from more or less systematically argued views of the proper structure and ends of human society. The history and general nature of ideological thought is considered in some detail in the *Macropædia* article IDEOLOGY. This article describes the major categories of modern ideology. Individual ideologies associated with particular theorists or historical settings are discussed separately elsewhere in the encyclopaedia; *e.g.*, for a full treatment of Marxism, see the *Macropædia* article MARXISM; for information on Nazism, see in the *Micropædia* under NATIONAL SOCIALISM and NAZI PARTY.

This article is divided into the following sections:

Socialism 393
- Origins of the socialist idea 394
- Marx and the rise of social democracy 394
- Other socialist tendencies before World War I 396
- The rise of Russian socialism 397
- Socialism between the wars 398
- Socialism after World War II 399

Communism 402
- The origins of Soviet Communism 402
- The Third International 402
- Stalinism 403
- Growth of Communism during and after World War II 404
- The world movement up to Stalin's death 405
- The breakup of the world Communist monolith 406
- Problems of internal reform 407
- Communist doctrine after Stalin 408

Anarchism 409
- Anarchist thinkers 409
- Anarchism as a movement 410
- Contemporary anarchism 413

Fascism 413
- National fascisms 413
- Common characteristics of fascism 414
- Varieties of fascism 415
- Origins and support 416
- Neofascism 417

Nationalism 419
- European nationalism 420
- Asian and African nationalism 421

Liberalism 422
- Classical liberalism 423
- Liberalism in the 19th century 425
- Modern liberalism 425
- Contemporary liberalism 426
- Conclusion 427

Conservatism 427
- Conservative attitudes 427
- Varieties of conservatism 428
- Modern conservatism 431
- Conservatism at the turn of the 21st century 433

Bibliography 433

Socialism

Socialism refers to both a set of doctrines and the political movements that aspire to put these doctrines into practice. Although doctrinal aspects loomed largest in the early history of socialism, in its later history the movements have predominated over doctrine, so much so that there is no precise canon on which the various adherents of contemporary socialist movements agree. The most that can be said is that socialism is, in the words of Anthony Crosland, a British socialist, "a set of values, or aspirations, which socialists wish to see embodied in the organization of society."

Problems of definition

Although it is possible to trace adumbrations of modern socialist ideas as far back as Plato's *Republic*, Thomas More's *Utopia*, and the profuse Utopian literature of the 18th-century Enlightenment, realistically, modern socialism had its roots in the reflections of various writers who opposed the social and economic relations and dislocations brought by the Industrial Revolution. They criticized what they conceived to be the injustice, the inequalities, the suffering brought about by the capitalist mode of production and the free and uncontrolled market on which it rested. To the acquisitive individualism of their age they opposed a vision of a new community of producers bound to each other through fraternal solidarity. They conceived of a future in which the masses would wrest control of the means of production and the levers of government from the capitalists.

Although the great majority of men calling themselves socialists in the 19th and 20th centuries have shared this vision, they have disagreed about its more specific ideas. Some of them have argued that only the complete nationalization of the means of production would suffice to implement their aims. Others have proposed selective nationalization of key industries, with controlled private ownership of the remainder. Some socialists insist that only strong centralized state direction and a command economy will suffice. Others advocate a "market socialism" in which the market economy would be directed and guided by socialist planners.

Socialists have also disagreed as to the best way of running the good society. Some envisage direction by the government. Others advocate as much dispersion and decentralization as possible through the delegation of decision-making authority to public boards, quasi-public trusts, municipalities, or self-governing communities of producers. Some advocate workers' control; others would rely on governmental planning boards. Although all socialists want to bring about a more equal distribution of national income, some hope for an absolute equality of income, whereas others aim only at ensuring an adequate income for all, while allowing different occupations to be paid at different rates.

"To each according to his need" has been a frequent battle cry of socialists, but many of them would in fact settle for a society in which each would be paid in accordance with his contribution to the commonwealth, provided that society would first assure all citizens minimum levels of housing, clothing, and nourishment as well as free access to essential services such as education, health, transportation, and recreation.

Socialists also proclaim the need for more equal political rights for all citizens, and for a levelling of status differences. They disagree, however, on whether difference of status ought to be eradicated entirely, or whether, in practice, some inequality in decision-making powers might not be permitted to persist in a socialist commonwealth.

The uses and abuses of the word socialism are legion. As early as 1845, Friedrich Engels complained that the socialism of many Germans was "vague, undefined, and undefinable." Since Engels' day the term socialism has been the property of anyone who wished to use it. The same Bismarck who as German chancellor in the late

1870s outlawed any organization that advocated socialism in Germany declared a few years later that "the state must introduce even more socialism in our Reich." Modern sophisticated conservatives, as well as Fascists and various totalitarian dictators, have often claimed that they were engaged in building socialism.

ORIGINS OF THE SOCIALIST IDEA

The utopians of the early 19th century

The term socialism, in its modern sense, made its first appearance around 1830. In France it was applied to the writings of Fourier and the Saint-Simonians and in Britain to those of Robert Owen.

Saint-Simon and Fourier. Comte Henri de Saint-Simon (1760–1825) was an erratic genius with a fertile and yet disorganized mind. His socialist writings revolved around the idea that his age suffered from an unhealthy and unbridled individualism resulting from a breakdown of order and hierarchy. But he held that the age also contained the seeds of its own salvation, which were to be found in the rising level of science and technology and in the industrialists and technicians who had already begun to build a new industrial order. The joining of scientific and technological knowledge to industrialism would inaugurate the rule of experts. The new society could not be equalitarian, Saint-Simon argued, because men were not equally endowed by nature. Yet it would make the maximum use of potential abilities by assuring that everyone would have equal opportunity to rise to a social position commensurate with his talents. By eradicating the sources of public disorder, it would make possible the virtual elimination of the state as a coercive institution. The future society would be run like a gigantic workshop, in which rule over men would be replaced by the administration of things.

Saint-Simon's followers bent the founder's doctrine in a more definitely socialist direction. They came to see private property as incompatible with the new industrial system. The hereditary transmission of power and property, they argued, was inimical to the rational ordering of society. The rather bizarre attempt of Saint-Simon's followers to create a Saint-Simonian church should not obscure the fact that they were among the first to proclaim that bourgeois-capitalist property was no longer sacrosanct.

François-Marie-Charles Fourier (1772–1837), a lonely and neglected thinker who was more than a little mad, was led to his anticapitalist vision by a loathing for a world of competition and wasteful commerce in which he spent most of his life as a salesman. Possessed by an inordinately wide-ranging imagination, he argued that the regenerated world to come would be characterized not only by social but also by natural and even cosmological transformations. The ocean would be changed into lemonade, and wild animals would turn into anti-lions and anti-tigers serving mankind.

Fourierist communities

With meticulous and obsessive care, Fourier set forth plans for his model communities, the *phalanstères*, the germ cells of the good society of the future. In these communities men would no longer be forced to perform uncongenial tasks but would work in tune with their temperaments and inclinations. They would cultivate cabbages in the morning and sing in the opera in the evening. Fourier's was an antinomian vision in which human spontaneity made outside regulation unnecessary. Whereas Saint-Simon called for the rule of experts, Fourier was convinced that love and passion would bind men together in a harmonious and noncoercive order.

Owenism. The Welshman Robert Owen (1771–1858) held more sober views. Early in his career he became known as a model employer in his textile works in Scotland, and as an educational and factory reformer. Despairing of his fellow capitalists he later turned to the emergent trade union movement. Acutely conscious of the evils of industrialism by which he had acquired his wealth, he thought that the new productive forces could be turned to the benefit of mankind if competition were eliminated and the effects of bad education were counteracted by rational enlightenment. He advocated cooperative control of industry and the creation of Villages of Unity and Cooperation in which the settlers, in addition to raising crops, would improve their physiques as well as their minds. Owenite communities established in New Harmony, Indiana, and elsewhere in America all failed. His attempts to join the cooperative and the trade union movements in a "great trades union" also proved a failure. Yet he left a lasting imprint on the British socialist tradition; his indictment of the competitive order, his stress on cooperation and education, his optimistic message that men could increase their stature if only the stultifying effects of an unhealthy environment were removed have continued to inform the socialist movement.

Other early socialists. The 1840s saw the rise of a number of other socialist doctrines, particularly in France. Louis-Auguste Blanqui evolved a radical socialist—or, as he called it, communist—doctrine based on a democratic populism and on the belief that capitalism as an inherently unstable order would soon be replaced by cooperative associations. Impatient with theorizing, given to a strong belief in voluntarism and the virtues of revolutionary action, he is remembered for his many attempts at organizing insurrections rather than for his theoretical contributions.

Étienne Cabet, in his influential utopian work *Voyage en Icarie* (1840), carried on the tradition of Thomas More as well as of Fourier. Louis Blanc is best known for *L'Organisation du travail* (1839), in which he advocated the establishment of national workshops with capital advanced by the government. These workshops would remain free from government control, with workers electing their management. The national workshops he organized in Paris after the revolution of 1848 were soon dissolved by a resurgent middle class. His plans for the "organization of labour" and his pleas for the recognition of the "right to work" were nevertheless a foreshadowing of the modern welfare state.

Pierre-Joseph Proudhon (1809–65) is best viewed as one of the founders of the anarchist tradition. But his attacks against private property and the institutions on which it rests, as well as his championing of a system of human relationships in which reciprocity, equity, and justice would replace what he saw to be rapacity, exploitation, and greed, powerfully stimulated the socialist imagination. His anti-statist and federalist vision of producers' communities provided a counterweight to the centralizing and statist impulses in the socialist tradition.

In England, the first half of the 19th century saw the emergence of a number of writers attacking the inequities of capitalism and basing their indictment of wage labour on radical interpretations of the thinking of an eminent economist, David Ricardo. Somewhat later, a Christian socialist movement led by Frederick Denison Maurice and Charles Kingsley attempted to combine radical economic views with political conservatism. The radical Chartist Movement of the 1830s and 1840s is better viewed as a political movement of the working class than as a specifically socialist formation, though anticapitalist ideas played a strong part in it.

MARX AND THE RISE OF SOCIAL DEMOCRACY

The idea of class struggle

In the perspective of intellectual history, all of these pre-Marxist socialist thinkers produced ideas of considerable intrinsic worth. But from the viewpoint of the subsequent development of socialism their ideas seem to be tributaries feeding the mighty stream of the Marxist movement that came to dominate the socialist tradition in the last third of the 19th century.

The Communist Manifesto. Karl Marx (1818–83) had a synthesizing mind. He fused German idealistic philosophy with British political economy and French socialism. Marx's earlier writings are discussed elsewhere (see MARXISM). In this section the focus is on his mature thought as first developed in *The Communist Manifesto* (1848), which he wrote in conjunction with Friedrich Engels, his lifelong intellectual companion.

To Marx, society is a moving balance of antithetical forces; strife is the father of all things, and social conflict is the core of the historical process. Men struggle against nature to wrest a livelihood from her. In the process they enter into relations with one another, and these relations differ according to the stage they have reached in their productive activities. As a division of labour emerges in

human society, it leads to the formation of antagonistic classes that are the prime actors in the historical drama. In contrast to his predecessors, Marx did not see history as simply a struggle between the rich and the poor, or the powerful and the powerless; he taught that such struggles differ qualitatively depending on what particular historical classes emerge at a given stage in history. A class is defined by Marx as a grouping of men who share a common position in the productive process and develop a common outlook and a realization of their mutual interest.

Marx, like Hegel and Montesquieu, considered societies as structured wholes; all aspects of a society—its legal code, its system of education, its religion, its art—are related with one another and with the mode of economic production. But he differed from other thinkers in emphasizing that the mode of production was, in the last analysis, the decisive factor in the movement of history. The relations of production, he held, constitute the foundation upon which is erected the whole cultural superstructure of society.

Marx distinguished this doctrine, which he called scientific socialism, from that of his predecessors whom he labelled utopian socialists. He asserted that his teachings were based on a scientific examination of the movement of history and the workings of contemporary capitalism rather than simply on idealistic striving for human betterment. He claimed to have provided a guide to past history as well as a scientific prediction of the future. History was shaped by class struggles; the struggle of contemporary proletarians against their capitalist taskmasters would eventuate in a socialist society in which associated producers would mold their collective destinies cooperatively, free from economic and social constraints. The class struggle would thus come to an end.

The First International. *The Communist Manifesto,* which had been written as a program for the Communist League, a group of continental workmen, failed to have an impact on the European revolutions of 1848. For a number of years thereafter Marx and Engels lived in complete isolation from the labour movements developing in England and on the Continent. Socialism in those years was only the creed of isolated sects, often of exiles. In 1864, however, after a gathering in London of continental and English workers' representatives and associated intellectuals, there emerged the International Working Men's Association, commonly known as the First International. Although it encompassed various tendencies ranging from simple trade unionism to anarchism, Marx dominated it from its inception and made it an instrument for the diffusion of his message. Its headquarters were in London, but it never exerted much influence in England, where the labour movement remained impervious to Marxist revolutionary ideology. On the Continent, particularly in Germany, Marxism spread rapidly and soon became the major doctrine of the emerging labour movement.

The growth of organized parties

German Social Democracy. In Germany, Ferdinand Lassalle (1825–64), the architect of the German labour movement, agreed with Marx on the need for autonomous organization of the working class but differed from him in wanting the government to provide the necessary capital for the establishment of producers' cooperatives that would emancipate labour from capitalist domination. To Marx, any appeal to the bourgeois state was out of the question, and he proceeded to organize followers in Germany against Lassalle. In 1869 they created the Social Democratic Party. The division between the followers of Lassalle and those of Marx persisted until 1875, when the two parties united on the basis of a compromise program (which Marx sharply criticized for its Lassallean vestiges).

The German Social Democratic movement grew rapidly, despite Chancellor Otto von Bismarck's attempts to suppress it through anti-socialist legislation and to undercut its appeal by social reforms. In 1877 the Socialists obtained half a million votes and a dozen members in the Reichstag. In 1881 the party claimed 312,000 members, and, by 1890, 1,427,000. After the repeal of the anti-socialist laws the party adopted the so-called Erfurt Program of 1891, eliminating all demands for Lassallean state-aided enterprises and pledging itself to the orthodox Marxian goal of "the abolition of class rule and of classes themselves."

Revisionism. It soon became apparent that Marx's own thought had gone through a process of evolution so that different disciples could quote chapter and verse in support of fairly divergent political views. In particular, whereas Marx in the late 1840s and early 1850s had asserted that only a violent revolutionary overthrow of bourgeois rule and the emergence of the "dictatorship of the proletariat" would lead to the emancipation of the working class, by the late 1860s his views had considerably mellowed. Writing in England after the second Reform Bill (1867), which had given the vote to the upper strata of the workers, Marx suggested the possibility of a peaceful British evolution toward socialism. He also thought that such a peaceful road might be possible in the United States and in a number of other countries.

Although the leaders of German Social Democracy liked to speak in revolutionary Marxist rhetoric, they had in daily life become increasingly absorbed in parliamentary activities. Under the intellectual guidance of their theoretician Karl Kautsky (1854–1938) they developed a brand of economic determinism according to which the inevitable development of economic forces would necessarily lead to the emergence of socialism. The official Social Democratic platform remained ideologically intransigent, while the party's activities became increasingly pragmatic.

Eduard Bernstein (1850–1932), once a close companion of Engels, challenging prevailing orthodoxy in his famous *Die Voraussetzungen des Sozialismus und die Aufgaben der Sozialdemokratie* (1899; Eng. trans., *Evolutionary Socialism,* 1909), appealed to the party to drop its revolutionary baggage and recognize theoretically what it had already accepted in practice: namely, that Germany would not have to go through revolutionary convulsions in order to reach socialist goals. Ignoring the differences between political conditions in Germany and England, Bernstein urged the party to travel along the English road in hope of gradually transforming capitalism through socialist reforms brought about by parliamentary pressure.

The struggle between Kautsky's orthodoxy and Bernstein's revisionism shook the German party. Bernsteinian doctrine was officially defeated in 1903, but revisionism in fact permeated the party, especially its parliamentary and trade union leaders. At the outbreak of World War I practically all the leaders supported the government and the war, thus ending the party's revolutionary pretensions.

Defeat of revisionism

Other Social Democratic parties on the Continent. In France, the Marxists had to contend with rival socialist traditions that had profound roots in French working-class history. The followers of Blanqui and Proudhon played leading roles in the Paris Commune of 1871. In the years that followed, French socialism was torn by conflicting tendencies. The Parti Ouvrier founded by Jules Guesde in 1875–76 represented Marxist orthodoxy, but there were other socialist parties that reflected the influence of Blanqui, Blanc, and Proudhon, as well as the 18th-century revolutionary heritage. Even after the various parties amalgamated in 1905, the movement continued to be torn by dissension between its revolutionary and reformist wings. Nonetheless, it continued to grow. At its first congress the unified party claimed 35,000 members, and in the elections of 1906 it won 54 seats in Parliament. By 1914 it had more than 100 members in the Chamber of Deputies. As in Germany, however, revolutionary rhetoric usually went hand in hand with pragmatic action, and the party became in fact a skillful participant in the parliamentary games of the Third Republic. After Jean Jaurès, the great Socialist orator and a principled leader of the peace elements, was assassinated on the eve of World War I, most of the Socialists supported the French war effort.

In the last part of the 19th century, Social Democratic parties generally beholden to Marxist doctrine sprang up in most of the countries of continental Europe. A Danish Social Democratic Party was founded in the 1870s, the Swedish Socialist movement in 1889. The Norwegian Labour Party (first called the Social Democratic Party) was formed in 1887 but became a major political force only in the early 20th century. In central Europe, Social Democratic parties fairly rapidly assumed a major place on the political horizon. An Austrian Social Democratic

Party was founded in 1888. By 1908 it had gained about one-third of the vote cast in the parliamentary elections, to become the strongest Socialist party outside Germany. The Belgian Labour Party, formed in 1885 as an amalgamation of trade union, cooperative, and other groups, rapidly organized thousands of mutual aid societies, built a very strong trade union movement, and led a number of general strikes on behalf of more liberal suffrage laws. The Dutch Socialist-Democratic Workers Party, founded in 1894, became a significant force only in the years immediately preceding World War I. It held 20 percent of the seats in the lower house of Parliament in 1912.

All of the continental parties were torn by internal tensions. Proposals to enter liberal coalition governments often were defeated by only narrow margins; Marxist orthodoxy prevailed only after sharp struggles. In The Netherlands, for example, a proposal to enter a coalition government was rejected by the close vote of 375 to 320 at the party congress of 1913.

Anarchist tendencies. In the less industrialized parts of Europe, particularly in Italy and Spain, Marxism had to contend with anarchist tendencies mainly rooted in the precapitalist and peasant strata. European anarchism as a political force was created by Mikhail Bakunin, the highly influential Russian libertarian thinker. His Anarchist Federation had belonged to the First International, but quarrels with Marx led to the expulsion of Bakunin and his followers in 1872.

The Anarchist Federation

Bakuninist and other anarchist strains of thought remained powerful in Spain, despite the founding of the Social Labour Party in 1879. The Spanish socialist movement suffered from the competition of the anarchists throughout its subsequent history, and only after World War I did it become a political force to be reckoned with.

In Italy anarchist tendencies also impeded the growth of a socialist movement. The Italian representatives to the First International followed Bakunin's lead. Not until 1892 was a distinctly Socialist party formed under the leadership of Filippo Turati. In 1913, after the electoral franchise was broadened, the official Socialist Party secured 51 seats in Parliament, and two other Socialist parties that had split from its ranks gained 31 seats. Although it continued to suffer from internal dissension and from anarchist tendencies in the more backward areas of the country, by World War I the Italian Socialist Party had become one of the strongest Marxist organizations in Europe.

The Second International. The First International had brought into being a variety of Socialist movements throughout Europe. When these began to grow roots in their respective political systems, it became apparent that the international movement could no longer be controlled by a single directing centre. After the dissolution of the First International in 1876, Marx and Engels remained father figures whose counsel the movement eagerly sought; but they could no longer direct it. The history of socialism now became largely the history of separate national movements that, for all their ceremonial acknowledgment of Marxist orthodoxy, increasingly tended toward a revisionist and nonrevolutionary line. By the early years of the 20th century socialism had become a powerful parliamentary force in most European countries. Except in Russia, where autocracy still held sway, the Socialists were reformers seeking a transformation of the existing system rather than its violent overthrow. Only left-wing minorities within the various parties still stood for revolutionary orthodoxy.

The Second International, founded in 1889, reflected the changed character of the movement. It was a kind of international parliament of socialist movements rather than the unified and doctrinally pure organization that the First International had attempted to be. It was dominated by the German party. With traditional Marxist rhetoric, the German delegates stood adamant against proposals to sanction socialist participation in bourgeois governments, and thus appeared to favour a "left" course. But socialist participation in government was not a realistic option in Kaiser William's Germany, and so the German delegates could be intransigent at no cost to themselves. When the issue was put to a vote at the Amsterdam congress in 1904, the Germans sided with those who opposed participation, against Jaurès and those who condoned it. But Jaurès had the better of it when he pointed out that "behind the inflexibility of theoretical formulas which your excellent comrade Kautsky will supply you with till the end of his days, you concealed . . . your inability to act." As with the issue of government participation, so with the issue of war. The Second International, under its German leadership, issued many moving and stirring manifestoes against war, but when war broke out it disclosed its paralysis. Most of its national components sided with their own governments and abandoned the idea of international working-class solidarity. Almost all of them recognized what they may secretly have believed for a long time: the workers, after all, had a fatherland.

OTHER SOCIALIST TENDENCIES BEFORE WORLD WAR I

British Fabianism. Although Marxism triumphed in the continental Socialist movement, it did not do so in Great Britain. Henry Hyndman, a radical journalist, founded the Social Democratic Federation on strictly Marxist principles in the 1880s, but it ever remained marginal to the British socialist movement. The Socialist League, founded by the poet William Morris, propounded libertarian-syndicalist ideas and likewise failed to make headway. Fabian socialism, on the other hand, based on non-Marxist ideas, was to have an enduring influence in Britain.

The non-Marxists

The Fabian Society was organized in the 1880s by a number of young radical intellectuals among whom Sidney and Beatrice Webb, Graham Wallas, Sidney Olivier, and George Bernard Shaw were the most outstanding. It developed an evolutionary and moderate form of socialism. Convinced of "the inevitability of gradualness," the Fabians never endeavoured to become a mass organization but preferred to be a ginger group of intellectuals working to transform society through practical and unobtrusive advice to the men of power. The extremely influential *Fabian Essays* (begun in 1889) contained detailed blueprints for social legislation and reform that influenced policymakers whether they were socialists or not. Through "permeation," which Shaw defined as "wire-pulling the government in order to get socialist measures passed," the Fabians attempted to convince key politicians, civil servants, trade union officials, and local decision makers of the need for planned and constructive reform legislation. Basing their doctrine at least as much on non-Marxist economics as on the continental socialist tradition, they worked for a new order "without breach of continuity or abrupt change of the entire social tissue."

Syndicalism. The syndicalist movement grew out of French trade unionism when it was reconstituted after the bloodletting of the Paris Commune (1871). Convinced of the futility of parliamentary and political activity, the syndicalists stressed that only direct action by workers organized in their unions would bring about the desired socialist transformation. Under the leadership of Fernand Pelloutier the Fédération des Bourses du Travail (founded in 1892), which was later amalgamated with the Confédération Générale du Travail (1902), was built on the idea that the emancipation of labour would come through a "general strike" that would paralyze the country and deliver power into the hands of organized workers. The unions would become the directing and administering nuclear cells of production.

The general strike tactic

The syndicalists attracted a number of intellectuals to their ranks, who attempted to provide a philosophical basis for syndicalism and its rejection of the political road to socialism. The most important of their writings, Georges Sorel's *Réflexions sur la violence* (1908; Eng. trans., *Reflections on Violence*, 1916), has continued to exercise considerable influence on the thinking of revolutionary militants, even though Sorel himself soon shifted his allegiance to the extreme right.

Guild socialism. The guild socialist tradition developed in Britain in the years before World War I. Sharing the general socialist hostility to the wage system and production for profit, guild socialists took from the syndicalists their distrust of the state and their emphasis on producers' control. They looked back to the Middle Ages when

independent producers, organized in guilds, controlled the conditions of their employment and took pride in creative work. Aiming at self-government in industry, guild socialists urged that industrial organizations, churches, trade unions, cooperative societies, and municipalities be granted autonomy. They argued that every group in society should carry out its particular functions without control from above, and that individuals should have a say in the direction of all those functional units in which they happened to be interested. Cooperation between functional units would replace direction by the state, which would be restricted to providing needed national services such as police protection. The state would be a functional unit among many others, rather than an all-encompassing sovereign.

Although guild socialism owes its origin to several thinkers, it grew into a mature doctrine only when in 1913 it recruited G.D.H. Cole, a brilliant Oxford don, two of whose early books, *The World of Labour* (1913) and *Self-Government in Industry* (1917), contain the best exposition of guild socialist doctrine. The movement never attained wide popular appeal but has continued to be a source of ideas in the British labour movement, if only as a counterpoint to the bureaucratic and centralizing tendencies of Fabianism.

Socialism in the United States. Socialism never became as influential in the United States as it did in Europe. When the Socialist Party was formed in 1901 it claimed a membership of 10,000 that grew to 150,000 in 1912, in which year the party polled a presidential vote of 897,000, or 6 percent of the national total. Although its strongest roots were among recent immigrants from Europe, it also drew its inspiration from the utopian colonies of the 19th century, from the slavery abolitionists, trade unionists, and agrarian reformers, and from isolated socialist groups of the 1880s and 1890s.

The Socialist Labor Party, a predecessor of the Socialist Party, was formed in 1877 but acquired a distinct outlook only when the journalist and polemicist Daniel De Leon joined it in 1890. De Leon attempted to marry a doctrinaire brand of Marxism to a "labourism" nourished in part on French syndicalist doctrine. He and his followers wished to raise the membership of the unions above "paltry routine business" and prepare them for a successful contest with the power of capital, both at the ballot box and in industrial combat.

The U.S. Socialist Party

The Socialist Labor Party remained a sect. But the Socialist Party developed into a mass movement under the leadership of Eugene Debs, a former union official who had been converted to socialism by reading the works of various socialist writers while in jail. The Socialist Party of Debs was neither centralized nor politically homogeneous. In its ranks it harboured reformists and revolutionaries, orthodox Marxists, Christian ministers, municipal reformers, populists who hated the railroads and the trusts, and Jewish garment workers dreaming of fraternity in the sweatshops. It produced no major theoretical works, but it managed in its undoctrinaire way to be an effective voice for the idea of socialism in America. It declined after World War I, its last well-known leader being Norman Thomas.

THE RISE OF RUSSIAN SOCIALISM

The populist tradition. The dominant radical tendency in 19th-century Russia was populism, a doctrine first developed by the author and editor Aleksandr Herzen, who saw in the peasant communes the embryo of a future socialist society and argued that Russian socialism might skip the stage of capitalism and build a cooperative commonwealth based on ancient peasant tradition. Herzen idealized the peasantry. His disciples inspired many students and intellectuals to "go to the people" in order to stir them into revolutionary action.

In the 1860s and 1870s, the more radical populists lost their faith in a peasant revolt and turned instead to terrorism. Small groups of student revolutionaries sought to bring down tsarism through terroristic action; their efforts culminated in the assassination of Alexander II in 1881. Sergey Nechayev's *Revolutionary Catechism*, in the writing of which Bakunin had a hand, stressed that the sole aim of the revolutionary is to destroy "every established object root and branch, [to] annihilate all state traditions, orders and classes in Russia." It is one of the ironies of history that Bakunin helped create in Russia an elitist and terrorist movement composed almost exclusively of alienated intellectuals, while in western Europe he appealed to skilled craftsmen and peasants and appeared to be the heir of Proudhon.

Within the broad stream of populism, terrorism was opposed by an evolutionary socialism that put its faith in peaceful propaganda and the education of the masses. While the elitists pursued their campaign of terror, the gradualists stuck to propaganda among the people.

Marxism in prerevolutionary Russia. The father of Russian Marxism was Georgy Plekhanov, who began his socialist career as a populist and was converted to Marxism when he settled in Geneva in 1880; in 1883 he founded the first Russian Marxist organization, the Osvobozhdenie Truda (Liberation of Labour Group). Plekhanov thought Russian socialism ought to be based primarily on the growing factory proletariat. Rejecting Herzen's idea that Russia was exceptional, he held that the revolution would be European in character and that Russia's place in it would be determined by its own labour movement. In a variety of books and pamphlets in the 1880s and 1890s, Plekhanov attacked the populists and argued that Marx had shown the objective historical necessity of socialism. The laws of social evolution could not be flouted. A bourgeois revolution in Russia was inevitable in the course of industrial development. The organized working class would know how to take advantage of the bourgeois revolution and push it forward.

The Leninists

Against this German brand of Marxism, Vladimir Ilich Ulyanov (1870–1924), later to be known by his party name of Lenin, argued for a more militant approach to revolution. In *What Is To Be Done?* (1902) he formulated his characteristic doctrine. Socialism would be achieved only when professional revolutionaries succeeded in mobilizing and energizing the masses of workers and peasants. Left to themselves, the workers would get no farther than a trade union consciousness. A militant, disciplined, uncompromising organization of revolutionaries was needed to propel the masses into action.

Lenin's followers parted company with the other Russian Marxists at the second congress of the (illegal) Russian Social Democratic Workers Party held in London in 1903. The anti-Leninist position was formulated by the leader of the more orthodox Marxists, L. Martov, when he declared, "In our eyes, the labour party is not limited to an organization of professional revolutionaries. It consists of them, plus the entire combination of the active, leading elements of the proletariat . . . ."

The two factions within the Russian Social Democratic movement at first cooperated and even held joint meetings; the final split came only in 1912. Individual leaders switched from one faction to another (Plekhanov, who originally sided with Lenin, joined his opponents in 1904). Others, such as Leon Trotsky, attempted for a time to stay free from factional alignments. These disputes were fought out in the West, where most of the leaders of both sides lived as émigrés. Within Russia itself, however, Lenin's opponents (the Mensheviks) mainly attracted the better educated and skilled workers, as well as the Jewish intelligentsia, while his Bolsheviks tended to be most successful among the more backward strata of the working class.

Mensheviks and Bolsheviks

After the February Revolution of 1917 toppled the tsarist regime and installed a liberal and vaguely socialistic leadership, the Bolsheviks managed to extend their organization among the urban masses. When Lenin returned from exile in April 1917, he startled his followers by calling for an entirely new strategy. Previously they had believed that their immediate task was to work within the limits of a democratic republic while preparing for future revolutionary opportunities. Lenin argued instead that they must seek power at once. The desire of the masses for an immediate end to the war, the land hunger of the peasantry, the feebleness of the new regime, he urged, made possible what had not been possible in the abortive

revolution of 1905: a socialist revolution led by Bolshevik cadres. Moreover, Lenin argued, a Russian revolution would not be isolated for it would soon be followed by a German revolution.

The soviets (workers' and peasants' councils), which had sprung up spontaneously when the tsarist power collapsed, were the main organizational bases from which the Bolsheviks mounted their assault on the established order. Lenin's slogan "All power to the soviets" found a ready response in the major urban centres. In September 1917 the Bolsheviks won elections for the Moscow and St. Petersburg soviets. These now became centres of "dual power" challenging the official government. It was the St. Petersburg soviet that in October 1917 gave Trotsky the military instrument with which he was able to topple the provisional government and install a revolutionary regime headed by Lenin.

Lenin and the Third International. The Bolshevik seizure of power had been undertaken in the belief that the revolution would soon spread to the rest of Europe. Lenin's perspective had always been internationalist. When most of the socialist leaders of the Second International rallied to their national governments in 1914, Lenin denounced them as traitors to the cause and sought to lay the groundwork for a new organization of revolutionary Socialists. After their seizure of power, the Bolsheviks resolved to create a Third International. By the time the delegates had assembled in Moscow in 1919, a revolutionary uprising in Berlin had been crushed and its leaders murdered. The great majority of the German working class was evidently willing to give the Social Democratic leadership of the new German republic a chance. But to the Russian leaders' world revolution still seemed near. Soon after the first congress of the Third International a short-lived soviet republic was proclaimed in Hungary and another in the German state of Bavaria. Communist parties began to be organized in all the major countries of Europe.

When the so-called Communist International (Comintern) met for its second world congress in July 1920, it was no longer a small gathering of individuals or representatives of small sects but a union of delegations from a dozen major Communist parties. The outcome of this meeting was to give the Russian leaders control of the new International, now broken away sharply from the Socialist movement. It adopted 21 conditions for membership in the Comintern, demanding that its adherents reject not only those Socialist leaders who had been "social patriots" in the war but also those who had taken a middle position. It aimed at creating a disciplined and militantly revolutionary world organization patterned after the Russian model, which would accept willingly the direction and unquestioned authority of the Russian leadership.

By 1923 the hoped-for revolutionary tide in Europe had not developed. New uprisings in parts of Germany failed completely in 1923. The Red Army's attempted invasion of Poland had been thrown back. Many Socialists who had for a time joined the Comintern, including the leadership of the Norwegian Labour Party, left-wing Communists in Germany, and Syndicalists in France and Spain, now turned away, rejecting its policy of centralized dictation.

Europe achieved a measure of economic and social stabilization. By the time of Lenin's death in 1924, Moscow was beginning to use the parties over which it still held command as instrumentalities of Russian foreign policy. Although some Comintern leaders like Trotsky still believed that world revolution was on the agenda, their faith was no longer shared by the majority of the Russian leadership.

SOCIALISM BETWEEN THE WARS

Socialists and Communists

The split with the Communists. Communists throughout the world denounced the leaders of the reconstructed Socialist parties as "social traitors" who "objectively" fostered the maintenance of capitalism. They accused them of having repudiated Marxism and betrayed international socialism by collaborating during the war with the bourgeoisie in the defense of their national states. The Socialist leaders retorted by pointing to the dictatorial features of the Soviet state and accusing the Communists of having betrayed the democratic socialist tradition.

The European Socialist movement was irremediably split. In Germany, the Social Democrats united again and succeeded in enrolling the bulk of the working class under their banner; the Communists were reduced to a minority position in the German labour movement. In France, where the Communists at first succeeded in attracting the majority of the Socialist Party, their opponents soon regained ascendancy and the Communists became a minority on the French left. Italian socialism split into Communists and left-wing and right-wing Socialists and thus greatly facilitated Mussolini's march to power. In Great Britain the Communists hardly made a dent in the Labour Party and never became more than a radical sect. European socialism as a whole, as well as socialist movements on other continents, was sharply split between adherents of the Second International and the Communists organized in the Third.

The Comintern followed an erratic course, sometimes veering toward a revolutionary line and sometimes making attempts to collaborate with the more militant strata of the socialists. After the onset of the economic depression in 1929 the Comintern took a sharp leftist turn, expecting the "final crisis" of capitalism to bring proletarian revolution everywhere. It denounced Social Democratic leaders as "social Fascists" and enemies of the working class. In the Prussian Landtag the Communists actually voted with the Nazis to bring down a Social Democratic government, on the theory that the Nazi movement was a passing phenomenon.

At the same time the Socialists gave up in practice, though not always in theory, their commitment to revolutionary doctrine. They became in effect pressure groups trying to extract maximum advantages for the working classes from their respective national regimes. In Germany, in Britain, and in the Scandinavian countries they participated at times in the government. Elsewhere, as in France, they tended to support congenial left-bourgeois regimes. But they lacked, on the whole, a concrete plan of social and economic action, and consequently were ineffective when the world depression unsettled the economies and political regimes of western and central Europe.

Response to the world economic crisis. Nowhere, except in Sweden and Belgium, did the socialists press for comprehensive socialist planning during the depression. Where they were in power they followed orthodox policies of budgetary management and public finance. When they were out of power they contented themselves with a defense of the immediate interests of the workers by demanding more unemployment insurance and opposing reductions in wages.

As the crisis deepened, the Communists gained influence, particularly among the unemployed and those unskilled workers hit most severely by the depression. They did not make deep inroads among other workers.

The rise of Fascism. Hitler's rise in Germany led to the destruction of both the Communists and the Socialists in that country. The Communists had hoped that a Nazi victory would be only temporary, and that afterward they would be called upon to lead the masses of Germany to victory. Their battle cry was, "After the Nazis—We." The Socialists played politics as usual, expecting that the depression would run its "natural" course and that a gradual decline of the Nazi fever would follow. A disunited labour movement proved unable to stay the Nazi march to power. This disaster led both Communists and Socialists to reconsider their previous policies and to revise their strategy and tactics.

Austrian Socialists, threatened with destruction by the reactionary regime of Chancellor Engelbert Dollfuss, resolved to offer armed resistance in February 1934. The Austrian party had long been regarded as a model for both its theoretical contributions and its concrete accomplishments. It enjoyed the nearly total support of the workers; 500,000 of Vienna's 2,000,000 inhabitants were dues-paying members. But the party was almost completely metropolitan and urban. Consequently the bloody battles of February 1934 remained localized in Vienna. The uprising was suppressed after four days, and the party had to go underground.

Socialists in office

Experience in government. *Germany.* The end of World War I had seen a somewhat reluctant Social Democratic Party installed in the seat of German government. Friedrich Ebert, the head of the party, became the first president of the new republic. But the Socialists were split internally. The "majority Socialists," the right wing of the party, wished to proceed in a cautious and pragmatic manner. The "independent Socialists," led by Kautsky and his former antagonist Bernstein, pressed for fundamental structural reforms. The extreme left, led by Rosa Luxemburg and Karl Liebknecht, wished to organize a revolutionary party and founded the Communist Party of Germany. When younger extremists, overruling Luxemburg and Liebknecht, organized a left-wing Putsch early in 1919, they were isolated and easily defeated by the government of the majority Socialists and its allies among right-wing officers. Luxemburg and Liebknecht were assassinated, and the remaining leaders took the group into the Comintern. Another left-wing and Communist putsch in Bavaria a few months later was also unsuccessful. In the early 1920s, the independents reunited with the majority Socialists.

In the first election to the new National Assembly in 1919 the majority Socialists obtained a plurality of the votes cast (39.3 percent), and the independent Socialists won another 8 percent. The Socialist government proclaimed the need for socialization of monopolistic industries and other radical measures. But after the elections of June 1920, a non-Socialist cabinet took office. In subsequent years the cabinets were largely non-Socialist in character, though Socialists participated in some of them. The middle classes were again in the saddle, and when President Ebert died in 1925 the conservative nationalist Hindenburg succeeded him. Throughout the turmoil of the first years of the Weimar Republic, the Social Democrats remained a bulwark of republican legality against both the extreme right and the extreme left. In the *Länder* (states), Prussia in particular, they held positions of governmental power and managed to institute a number of reformist welfare measures. But they failed to gain a controlling voice in national politics.

In the May 1928 elections the Social Democrats emerged as the strongest party in the Reichstag. Although they lacked a majority, their leader Hermann Mueller became chancellor, and their financial expert was named minister of finance. This largely Socialist government, however, proved unable to deal with the economic depression that soon afflicted Germany along with the rest of the world. The government followed an orthodox deflationary policy, pressed for the reduction of unemployment benefits in order to save taxes, and attempted to reduce budget deficits. Unable to stem the tide of depression, it resigned in 1930. This was the last government of the Weimar Republic in which Social Democrats participated. Soon afterward, the Nazis started on their way to power.

Britain. In the general election of 1923 the Labour Party, which had adopted a Socialist program only five years earlier, won a plurality; with the support of the Liberals it formed the first Labour government under Ramsay MacDonald in January 1924. Its tenure proved short. After implementing a few modest reform measures, it was ousted by an electorate which, partly because of manufactured fears of a "Bolshevist menace," turned sharply to the right in the elections of October 1924.

In June 1929 the Labour Party had its second chance. It won 288 out of 615 seats in the House of Commons and, with the support of the Liberals, formed the second Labour government, again under Ramsay MacDonald. But Labour, like the German Social Democrats, proved unable to deal with the depression, particularly with mounting unemployment. It was pledged to far-reaching social reforms that it was not prepared to carry out. The flight of capital from London assumed catastrophic proportions; business circles demanded a balanced budget and lower unemployment benefits. When MacDonald proposed to accede to some of these demands, the trade unions sharply opposed him. He then split the Labour government and formed a national coalition with the Conservatives and the Liberals. For the remainder of the 1930s the Labour Party was out of power.

Italy. In the Italian elections of 1919, the Socialists won 2,000,000 votes out of a total of 5,500,000. Italy seemed on the verge of revolution; large-scale strikes, mass demonstrations, factory occupations, and spontaneous expropriations of landed estates spread throughout the country. In August 1920 a revolutionary situation developed in the industrial north after a breakdown in wage negotiations; 500,000 workers occupied the factories, kept production going, and prepared for armed resistance. The far left called for an extension of the strike, but a divided Socialist leadership hesitated. The discouraged workers retreated. Mussolini's Blackshirts began breaking up working-class meetings. In 1921 the right-wing Socialists proposed that the party form a coalition government with the Liberals, but the left vetoed the idea. Mussolini's terror squads made further inroads in the large industrial centres. A general strike called by the trade unions proved a dismal failure. Soon afterward Mussolini made his March on Rome (October 1922) and was installed as premier. By 1926 parliamentary government had completely ended in Italy. The Socialists were driven underground.

Italian crises of 1919–22

France. None of the French governments from the end of World War I until the middle 1930s included Socialists. Although the Socialist Party was in fact deeply committed to gradualism, it still clung to its prewar policy of not participating in "bourgeois" governments. Only in the mid-1930s, when militant right-wing groups threatened the Third Republic, did the Socialists change their policy. In June 1936 a government took office representing a Popular Front, ranging from the Communists on the left to Radical Socialists in the centre and headed by the Socialist leader Léon Blum. The Communists had at last abandoned their doctrine of "social Fascism" and were now willing to enter coalitions with other parties of the centre and left.

The victory of the Popular Front in June 1936 was accompanied by sit-down strikes in the factories; these helped push the government, headed by Léon Blum, in a radical direction. Collective bargaining rights, never recognized before by French employers, were now protected by law; social security and general working conditions were significantly improved; the 40-hour week was made mandatory. The Blum government attempted to institute a French version of the U.S. New Deal. But after the initial enthusiasm had waned, French employers took courage and pressed the government to return to traditional fiscal and budgetary policies. When in June 1937 his middle-of-the-road partners in the coalition refused his demands for emergency fiscal powers, Blum resigned. The Socialists participated in the next government headed by a Radical Socialist, and Léon Blum later formed another Popular Front government that held office for about a month in 1938. When France went to war against Germany in 1939 the Communist Party, which opposed the war, was banned. After France's collapse in 1940, the Socialist Party was dissolved by the Vichy government.

Sweden. Only in Sweden were Socialists successful in their governmental policies. A Swedish Labour government was formed for the first time in 1932. Unlike the other European Socialist parties, the Swedes broke with orthodox budgetary and financial policies and stressed large-scale intervention by the government in the planning of economic affairs. Extensive public works, financed by borrowing from idle capital resources, helped to reduce unemployment and stimulated the economy; public investment was used methodically to offset the effects of reduced private spending. Unemployment, which had reached 164,000 in 1933, was eliminated by 1938 through a policy of steady economic expansion. The Swedish innovations helped lead the way to the economic policies practiced by almost all Western countries after World War II.

SOCIALISM AFTER WORLD WAR II

The worldwide spread of "socialist" parties. Orthodox Marxists had always assumed that socialism would emerge first in the industrial countries of the world. But a new kind of "socialism" spread rapidly in agrarian societies and backward countries after World War II. In many of these countries Marxism became, despite the intention of its founders, the ideology of industrialization. In the struggle against colonialism the liberation movements, especially the intellectuals and semi-intellectuals who led

New concepts of socialism

them, adopted what they conceived to be socialist ideas. It seemed to them that meaningful national independence could be attained only through state direction of the economy. Rapid economic growth, they believed, could be fostered only by restricting consumption and channelling national resources into the building up of productive facilities. In one degree or another the new countries took the Soviet Union as their model for rapid industrialization. All manner of regimes, from totalitarian one-party states to military dictatorships, proclaimed that they were socialist. Only in India and a very few other countries did the ruling party retain the traditional Western socialist vision of social justice, equality, and democracy.

In the meantime, ironically, the Socialists of western Europe were giving up their Marxist views and turning toward the welfare state. During World War II almost all of the Socialist parties had joined governments of national unity. Afterward they sought to become popular parties following the parliamentary road to power and ready to participate in coalition governments with Liberal or Christian Democratic partners. Surrendering the idea that only full state ownership would bring the good society, they aimed at a mixed economy in which public control and a certain amount of planning would bring social benefits for all. This was, in essence, the idea of "the inevitability of gradualism" that the Engish Fabians and the German revisionists had preached around the turn of the century.

From socialism to welfare

The transformation of western European socialism. *Germany.* The new orientation of the postwar Social Democratic Party of Germany was expressed in its Frankfurt declaration of 1951. It advocated public control of the economy but rejected comprehensive state ownership, and it stressed that democratic socialist planning had nothing in common with the Communist and totalitarian kind.

A few years later, in its program adopted in Bad Godesberg in 1959, the party shed the last remnants of Marxism. The name of Marx and the words "class" and "class struggle" did not appear in the program, which even advocated private property in the means of production. It rejected overall central planning and endorsed the idea of a competitive free market. The party now stood for "as much competition as possible—as much planning as necessary." Moreover, the party no longer claimed to possess a universally valid doctrine and instead embraced a pluralistic society in which no party would seek to impose its particular philosophy on society as a whole. Thus, the Social Democratic Party of Germany had become a reformist party striving for an extension of the welfare state.

With their election victory in 1969, the Social Democrats became the senior party in a coalition that governed West Germany for the next 20 years. Their efforts to extend the welfare state, however, were largely thwarted by anti-interventionist elements in their coalition partners, the Free Democrats. After a long interval of Christian Democratic rule beginning in 1982, the Social Democratic Party returned to power in 1998. Under its new leader, Gerhard Schröder, the party further moderated the social-welfare policies it had advocated in earlier years and even proposed curtailing some existing programs.

Britain. The British Labour Party was never committed to Marxism and hence found it easier to adjust to the political realities of the postwar world. In 1945 it won a majority in Parliament for the first time. The government of Prime Minister Clement Attlee, during its six years in power, laid the foundations of the British welfare state. A number of basic industries, such as coal, railways, road transport, and steel, were nationalized. A comprehensive system of nationalized medical care was established. Social services were extended. Full employment was maintained. Although Labour was voted out of office in 1951, its main achievements remained. The steel industry again reverted to private control, but for the next three decades no efforts were made to undo other features of the welfare state.

Hugh Gaitskell, who succeeded Attlee in the leadership of the party, wanted to revamp its program by eliminating earlier pledges that the party seek large-scale nationalization of industry. He was not successful, but in practice the party adopted a reformist course aimed at the extension of the welfare state and pragmatic planning. When the party returned to power in 1965, its leader, Harold Wilson, prime minister until 1970, pursued a cautiously reformist policy. Harassed by economic difficulties, the Labour government made few policy decisions of a distinctively socialist character.

The Labour Party returned to office in 1974, facing high inflation and considerable labour unrest. After its supporters in the trade unions paralyzed the country during the winter of 1978–79, the party fell from power and remained in opposition for the next 18 years, during which much of the British welfare state was dismantled by a resurgent Conservative Party under Margaret Thatcher. When "New Labour" came to power in 1997, it was not as a socialist party but as a party committed to managing a mixed economy more efficiently than its Conservative predecessors, though not in a very different fashion from them.

France. The French Socialist Party held leading positions in the first few postwar French governments. It supported nationalization of some parts of French industry, public control of the economy, and the reform of social security. But the party had lost much of its prewar support among the workers to the Communists, and increasingly it became a party of civil servants, middle-class professionals, and other white-collar employees. Although the Socialists made no attempt to recast their program as the German Social Democrats did, their policies were just as moderate. When party leader François Mitterrand became the first leftist president of France in 1981, the Socialists nationalized a number of industrial and financial concerns, but a worldwide recession and pressures on the franc eventually forced them to retreat from many of these initiatives. In 1997, Socialist leader Lionel Jospin became prime minister on a program promising a shortened workweek, reduced unemployment, and a more moderate approach toward France's entry into the European Union.

Socialist government in France

Italy. After World War II the Italian Socialist movement split into a number of parties. The largest, the Italian Socialist Party under Pietro Nenni, attempted to revive the left-wing socialist tradition of the pre-Mussolini era. In 1947 the party split over the issue of whether to cooperate with the Italian Communist Party, with a majority under Nenni favouring cooperation and a minority, under Giuseppe Saragat, rejecting it. Saragat's faction, which became the Italian Social Democratic Party (originally the Socialist Party of Italian Workers), was committed to moderate social reforms and participated in most centre-left Italian governments through the 1980s.

After the failed revolt against Communist rule in Hungary in 1956, the Italian Socialist Party increasingly distanced itself from the Communists, and eventually it joined a coalition government with the Christian Democrats (1963). The party was a member of most Italian governments during the 1970s and '80s. In 1983–87 the Socialists formed their own government under Bettino Craxi, who became Italy's first Socialist prime minister. Following charges of corruption against Craxi and other party members in the early 1990s, the party's fortunes waned, as did those of the Italian Social Democratic Party. In 1994 the two parties merged for the second time (an earlier merger in the late 1960s was short-lived) and adopted the name Italian Democratic Socialists.

African socialism. Socialist ideas were carried to North Africa mainly by French-educated African intellectuals; in addition, many French settlers, especially schoolteachers and civil servants, were Socialists or Communists. The various national liberation movements, especially in Tunisia and Algeria, linked the struggle against colonial domination with socialist ideas. When Algeria became independent, its first leader, Ahmed Ben Bella, surrounded himself with French advisers from various Marxist groups.

Collectivization of agriculture and self-management in industry stood high on the agenda of the Algerian national government. When these programs failed, Ben Bella was replaced by Col. Houari Boumedienne, who was pledged to continue "Algerian socialism" but settled for an economy based on state-directed enterprises and private landholdings. In fact, the country was run by a military dictatorship. In 1988–89, under Col. Chandli Benjadid, Algeria adopted constitutional amendments that reduced the

Algerian socialism

formal powers of the governing National Liberation Front and introduced elements of a multiparty system. However, the Front was unable to establish broad-based electoral support, and during most of the 1990s it was relegated to the opposition in governments directly or indirectly controlled by the military.

In Tunisia a one-party regime was installed after independence in 1956, and under its leader, Habib Bourguiba, it proceeded to nationalize major enterprises. The ruling Destourian Socialist Party (after 1988 the Democratic Constitutional Assembly), which remained the only legal party until 1981, was committed to modernization through planned economic development. Despite gradual electoral reforms from the 1980s, the few legal opposition parties continued to face systematic obstacles, and they did poorly in elections through the end of the 20th century.

Elsewhere in Africa, the ruling elites proclaimed their adherence to one or another version of "African socialism" while in fact being committed above all to rapid industrialization and modernization. Many African socialists stressed the need to build their ideology upon African traditions such as communal land ownership, the egalitarian practices of some tribal societies, and the network of reciprocities and obligations that once existed in tribal societies. By the end of the 20th century, however, the attempt to construct African socialism had largely been abandoned, as many countries were forced to adopt privatization programs and other "structural reforms" in order to secure badly needed loans and to attract foreign investment.

Arab socialism. The "socialist" movements of the Middle East have been led by European-educated intellectuals belonging to a new middle class of civil servants, army officers, and schoolteachers. Attempting to appeal to the Arab people as a whole, they have stood for modernization and for the brotherhood of all Arabs.

The most important of these movements was the Arab Socialist Party, usually called the Ba'th Party. Founded in Syria in 1943 and subsequently established in many other Arab countries, it called for the formation of a single Arab socialist nation, though in practice it did little to promote specifically socialist policies. The party was most successful in Syria, where Ba'thist president Hafez al-Assad ruled from 1970 until his death in 2000, and in Iraq, where the sole ruler of the country from 1979 was Pres. Saddam Hussein. Hussein maintained his grip on power despite his country's defeat in the 1991 Gulf War and the ruinous international sanctions that followed. Hussein's regime, like Assad's, was only nominally socialist. Indeed, for many observers it exemplified a new brand of Arab fascism.

When Gamal Abdel Nasser came to power in Egypt in 1952, his group of young army officers had little if any interest in socialism. Nasser was subsequently led to socialist ideas through his struggle against the domination of Egypt by foreign businesses. In 1962 the Arab Socialist Union was established as the only legal political party in Egypt, and in accordance with its program the country nationalized all large industrial and financial enterprises, expropriated large landholdings, and placed all other important sectors of the economy under state control. In 1976–77 other parties were granted legal recognition, and the Arab Socialist Union was soon re-formed as the centrist National Democratic Party. As the party of the government, it maintained a virtual monopoly of power through the end of the 1990s.

Failure of socialism to take root in Asia

Asian socialism. *Southeast Asia.* In the 1950s the governments of several countries of South and Southeast Asia—India, Burma, Ceylon, Indonesia, and Singapore—called themselves Socialist. Yet the Socialist parties in these countries soon lost all power and influence. The several Socialist organizations in India were dwarfed by the ruling Congress Party, which was striving to unite many divergent political and social tendencies within its ranks. The Burmese Socialist Party, though for many years a partner in coalition governments, was outlawed when Gen. Ne Win seized power in 1962. Similarly, the Indonesian Socialist Party was abolished by President Sukarno in 1960. In the subsequent decade, nearly all the Socialist parties of the region played no significant role in the political life of their countries.

As the influence of the European-style Socialist parties waned, various authoritarian regimes arose speaking in socialist accents. Suharto, who succeeded Sukarno as president of Indonesia in 1967, retained Sukarno's vaguely socialist ideology of Pancasila (Bahasa Indonesia: "Five Principles")—which included a commitment to "internationalism" and "guided democracy"—but in fact his regime, like Sukarno's, was little more than a corrupt personal dictatorship. The Burmese military dictatorship of Ne Win similarly proclaimed Burma a socialist state, and socialism was also the official program of SLORC (the State Law and Order Restoration Council), which came to power in another military coup in 1988.

Japan. Only in Japan, by far the most developed of Asian countries, did traditional socialist organizations become firmly established. The Socialist Party of Japan, formed in 1901, endured varying periods of repression and harassment until 1946, when it won more than 90 seats in the Diet to become Japan's third strongest party, and a year later its leader, Katayama Tetsu, became prime minister in a coalition government (1947–48). In the 1960s the party split to form the leftist Japan Socialist Party and the rightist Democratic Socialist Party. The Japan Socialist Party, which called for nonalignment and a democratic transition to socialism, gradually moderated its commitment to socialism in the 1980s and '90s. In 1993, as the Social Democratic Party of Japan, the party participated in the first government in four decades not headed by the Liberal Democratic Party. In 1994 the party's leader, Murayama Tomiichi, became the first socialist prime minister of Japan in nearly 50 years. Soon thereafter the party's electoral fortunes dramatically reversed, and by the end of the 1990s it had ceased to be a significant force in Japanese politics.

Success in Australia, New Zealand, Canada

Other countries and regions. *Australia and New Zealand.* Socialism has deep roots in the British Commonwealth countries of Australia, New Zealand, and Canada. The Australian Labor Party was formed in 1901, just as the Australian Commonwealth came into existence. Only three years later its leader, J.C. Watson, became the world's first Labor prime minister, and by 1915 Labor had headed three national governments. It subsequently led the country during several crucial events in its history, including the two world wars and the onset of the Great Depression. The party formed the government again in 1973–75 and 1983–96. At the end of the 1990s its program called for racial and gender equality, a nonaligned foreign policy, and an end to the constitutional role of the British monarchy.

A loose Liberal-Labour alliance dominated New Zealand politics between 1893 and 1906, but the New Zealand Labour Party did not emerge until 1916. It grew steadily, coming to power in 1935 for the first of several periods of varying duration. Despite its principled pledge to bring about Socialism, the New Zealand Labour Party, like its Australian counterpart, was committed to gradual reform, and it was mainly concerned with using governmental control as a means of dealing with immediate problems and expanding social services. In 1984, the Labour government of David Lange retreated from some of the party's earlier commitments by partially deregulating the economy and reducing some government subsidies. After seven years in opposition in the 1990s, the Labour Party returned to power in 1999 under its leftist leader Helen Clark.

Canada. Canadian socialism developed more slowly than its Australian and New Zealand counterparts. Prior to World War I the Canadian Socialist movement was split between two parties, neither of which managed to win seats in the federal Parliament. During the 1920s various Socialist and Labour parties flourished in different parts of Canada, but only rarely did they win federal office. In 1944 the socialist Co-operative Commonwealth Federation (CCF), campaigning on a promise of "social and economic planning on a bold and comprehensive scale," won provincial elections in Saskatchewan, and it remained in power there for the next 20 years. In 1961 progressive union leaders joined with the CCF to form the New Democratic Party, which expanded support for the Socialist movement from the agrarian midwest to the more industrialized parts of the country. Advocating a planned econ-

omy, the party stood for increased social security, government employment guarantees, and large-scale construction of low-rent housing, among other goals. The New Democrats formed governments in Saskatchewan, Manitoba, and British Columbia intermittently to the 1990s, in the Yukon Territory in the 1980s, and in Ontario, Canada's largest and richest province, in the 1990s.

Latin America. Socialism in Latin America has a long history. Several branches of the First International were established in Argentina in the early 1870s. In Chile and Argentina, and to a lesser extent in other countries of the region, socialists at times played leading roles, but they were hampered by factional conflicts and by the fact that their following consisted mainly of immigrant industrial workers. In Chile, however, they participated in coalition and popular-front governments in the 1920s, '30s, and '40s. In 1958, Chilean socialists supported the Popular Action Front (FRAP) candidate, Salvador Allende, who was narrowly defeated in that year and again in 1964. In 1970, Allende won by a narrow plurality in a three-way election and became head of a government supported by a broad popular front of leftist groups. Pledged to the nationalization of foreign-owned industry and to the planned reconstruction of the country, the government soon met with increasing economic and political turmoil, some of which was secretly abetted by the United States. Allende's death in a military coup in 1973 and 16 years of military dictatorship thereafter left the fate of socialism in Chile uncertain. However, the restoration of democracy led eventually to the election of a moderate socialist president, Ricardo Lagos, in 1999. (L.A.C./A.Ry./Ed.)

Chilean socialism

Communism

The word communism, a term of ancient origin, originally meant a system of society in which property was owned by the community and all citizens shared in the enjoyment of the common wealth, more or less according to their need. Many small communist communities have existed at one time or another, most of them on a religious basis, generally under the inspiration of a literal interpretation of Scripture. The "utopian" socialists of the 19th century also founded communities, though they replaced the religious emphasis with a rational and philanthropic idealism. Best known among them were Robert Owen, who founded New Harmony in Indiana (1825), and Charles Fourier, whose disciples organized other settlements in the United States such as Brook Farm (1841–47). In 1848 the word communism acquired a new meaning when it was used as identical with socialism by Karl Marx and Friedrich Engels in their famous *Communist Manifesto.* They, and later their followers, used the term to mean a late stage of socialism in which goods would become so abundant that they would be distributed on the basis of need rather than of endeavour. The Bolshevik wing of the Russian Social-Democratic Workers' Party, which took power in Russia in 1917, adopted the name All-Russian Communist Party in 1918, and some of its allied parties in other countries also adopted the term Communist. Consequently, the former Soviet Union and other states that were governed by Soviet-type parties were commonly referred to as "Communist" and their official doctrines were called "Communism," although in none of these countries had a communist society fully been established. The word communism is also applied to the doctrines of Communist parties operating within states where they are not in power. (For the ideological basis of Communism, see MARXISM, MARX AND.)

THE ORIGINS OF SOVIET COMMUNISM

Communism as it had evolved by 1917 was an amalgam of 19th-century European Marxism, indigenous Russian revolutionary tradition, and the organizational and revolutionary ideas of the Bolshevik leader Lenin. Marxism held that history was propelled by class struggles. Social classes were determined by their relationship to the means of production; feudal society, with its lords and vassals, had been succeeded in western Europe by bourgeois society with its capitalists and workers. But bourgeois society, according to Marxism, contained within itself the seeds of its own destruction: the number of capitalists would diminish, while the ranks of the impoverished proletariat would grow until finally there would be a breakdown and a Socialist revolution in which the overwhelming majority, the proletariat, would dispossess the small minority of capitalist exploiters.

Marxism had been known and studied in Russia for at least 30 years before Lenin took it up at the end of the 19th century. The first intellectual leader of the Russian Marxists was G.V. Plekhanov. Implicit in the teachings of Plekhanov was an acceptance of the fact that Russia had a long way to go before it would reach the stage at which a proletarian revolution could occur, and a preliminary stage would inevitably be a bourgeois democratic regime that would replace the autocratic system of Tsarism.

Plekhanov's Marxism

Plekhanov, like most of the early Russian Marxist leaders, had been reared in the traditional Russian revolutionary movement broadly known as Populism, a basic tenet of which was that the social revolution must be the work of the people themselves, and the task of the revolutionaries was only to prepare them for it. But there were more impatient elements within the movement, and it was under their influence that a group called "People's Will" broke off from the Populist organization "Land and Freedom" in 1879. Both groups were characterized by strict discipline and highly conspiratorial organization; "People's Will," however, refused to share the Populist aversion to political action, and in 1881 some of its members succeeded in assassinating Tsar Alexander II.

Lenin and Russian Populism. During the period of reaction and repression that followed, revolutionary activity virtually came to an end. By the time Lenin emerged into revolutionary life in Kazan at the age of 17, small revolutionary circles were beginning to form again. Lenin was a revolutionary in the Russian tradition for some time before he was converted to Marxism (through the study of the works of Marx) before he was yet 19. From the doctrines of the Populists, notably P.N. Tkachev, he drew the idea of a strictly disciplined, conspiratorial organization of full-time revolutionaries who would work among important sections of the population to win support for the seizure of power when the moment was ripe; this revolutionary organization would take over the state and use it to introduce Socialism. Lenin added two Marxist elements that were totally absent in Populist theory: the notion of the class struggle and the acceptance of the need for Russia to pass through a stage of capitalism.

Lenin's most distinctive contributions to Communist theory as formulated in *What Is to Be Done?* (1902) and the articles that preceded it were, first, that the workers have no revolutionary consciousness and that their spontaneous actions will lead only to "trade union" demands and not to revolution; second, the corollary that revolutionary consciousness must be brought to them from outside by their intellectual leaders; and third, the conviction that the party must consist of full-time, disciplined, centrally directed professionals, capable of acting as one man.

Lenin's concept of the revolutionary party

Lenin's tactics led in 1903 to a split in the Russian Social-Democratic Workers' Party. With his left-wing faction, called the Bolsheviks, he strove to build a disciplined party and to outwit and discredit his Social-Democratic opponents. After the collapse of tsarism in February 1917, he pursued a policy of radical opposition to the Socialists and Liberals who had come to power in the provisional government, and he eventually succeeded in seizing power in October 1917. Thereafter he eliminated both the opposition of other parties and his critics among the Bolsheviks, so that by the 10th party congress in March 1921 the Bolsheviks (or Communists) had become a monolithic, disciplined party controlling all aspects of Russian life. It was this machine that Stalin inherited when he became general secretary of the party in 1922.

THE THIRD INTERNATIONAL

The victory of the Bolsheviks in Russia gave a new impetus to the more extreme left wings of the Socialist parties in Europe. Lenin's relations with the European Socialist parties had been hostile even before World War I. During the war he had endeavoured to assert his influence over the dissident left wings of the Socialist parties of the bel-

ligerent powers, and at two conferences in Switzerland, in 1915 and in 1916, he had rallied these dissident groups to a policy of radical opposition to the war efforts of their governments and to an effort to turn the war into a civil war. He had already decided by 1914 that, after the war, a Third International must be formed to take the place of the Second International of Socialist parties, which had failed to oppose the war despite its strong antiwar tradition. By 1919, when the new Soviet regime in Russia was fighting for its survival, the intervention on the anti-Soviet side by Britain, France, and the U.S. was a powerful and practical argument to be used by Soviet Russia in its appeals for revolution in capitalist countries. It early became clear the Third International would reflect the influence of Soviet Russia and that it was likely to become subordinate to Soviet aims and needs.

Communism's emergence as an international movement

Lenin's 21 conditions. The Third International, or Comintern, had its first congress in 1919. This gathering of a very few parties in Moscow was more symbolic than real; the main structure of the new International was not hammered out until the second congress in July 1920, also in Moscow. Hopes of world revolution ran high; the prestige of the new Soviet state was in the ascendant, and the resolutions adopted at this congress reflected in the fullest possible way Lenin's idea of what a Communist party should be. It was to be the "main instrument for the liberation of the working class," highly centralized and disciplined according to the formula of "democratic centralism" on which the Bolshevik Party had been founded. Twenty-one conditions were laid down by the congress as prerequisites for parties affiliating with the Comintern. These conditions were designed to ensure a complete break with the older Social Democratic parties from which the Communist parties were splitting off. The new parties were required to adopt the name Communist in their title, to urge open and persistent warfare against reformist Social Democracy and the Second International, to maintain a centralized and disciplined party press, to conduct periodic purges of their ranks, and to carry on continuous and systematic propaganda in the army and among the workers and peasants. Each constituent party was to support in every possible way the struggle of "every Soviet republic" against counterrevolution. Decisions of the Comintern and of its executive committee were to be binding on all members, and the breach of any of these conditions was to be ground for expelling individual members from their parties—a provision that in future years was to be interpreted very broadly.

The New Economic Policy. The prestige of Soviet Russia, the rigid discipline imposed by the 21 conditions, and certain other factors ensured the predominance of Russian control and Russian interests over the Comintern. Though the predominance increased during Stalin's time, it was clearly evident while Lenin was still alive. At the third world congress in June and July 1921, the Comintern was confronted by Lenin with his New Economic Policy—a program encouraging small private enterprise, which several months earlier he had put into effect inside Russia. Lenin wanted a temporary halt to the revolutionary upsurge in Europe to give him time to develop stable trade relations with capitalist countries, to whom the Soviet state was preparing to grant trading and industrial concessions. Comintern members were required to support this policy, and the expulsion of the German Communist leader Paul Levi after the failure of a Communist uprising in Germany in March 1921 showed how determined the leaders of the Comintern were to put down inconvenient left-wing "adventures." It was with the requirements of the New Economic Policy in mind that the Comintern executive committee in December 1921 launched the turnaround policy of the United Front and of trade union unity. This policy of rapprochement with Socialists and liberals was likewise designed to gain support for Lenin's policy of consolidation at home by appealing to a broader spectrum of opinion in the capitalist countries.

STALINISM

Socialism in one country. Lenin's successor, Joseph Stalin, always claimed to be his faithful follower, and this was to some extent true. Stalin's doctrine that Socialism could be constructed in one country, the Soviet Union, without waiting for revolution to occur in the main capitalist countries (a position he had developed as an integral part of his struggle against Trotsky) was not far removed from the line pursued by Lenin in 1921 when he introduced the New Economic Policy. Both Lenin and Stalin accepted the primary importance of the survival and strengthening of the Soviet state as the main bastion of the future world revolution; both accepted the need for a period of coexistence and trade with the capitalist countries as a means of strengthening socialism in Soviet Russia. Nor did Stalin's later policy of industrialization and collectivization, in theory at least, represent a departure from Lenin's doctrine. Industrialization was central to Lenin's plans, though he did not live to put them into practice. Stalin's view, however, that the construction of socialism led inevitably to an intensification of the class struggle, which in turn required a policy of internal repression and terror, is nowhere to be found in Lenin's writings. On the contrary, Lenin repeatedly emphasized in 1922 and 1923 the necessity of bringing about a reconciliation of the classes and especially of the peasants and workers.

From Lenin to Stalin

Stalin's internal policy was to have wide repercussions in the Comintern and on Communism generally. From 1924 until 1928 his first concern was to defeat his main rival, Trotsky, and this seems to have been one of the main factors determining his policy at this time. As against the more internationalist and doctrinaire Trotsky, Stalin pursued "socialism in one country" and continued to implement Lenin's New Economic Policy with its limited freedom for business enterprise and peasant individualism. In this he could still claim to be following Lenin's wishes. But Stalin also worked with great skill to ensure his control over the party. By 1927 when Trotsky was expelled from the party, Stalin already controlled both the network of party officials (the *apparat*) and the delegates to congresses and conferences. Debate had been replaced by ritualized unanimity; dissent was permitted only when it served the purposes of the leadership.

When Trotsky was exiled from the country in 1929, he became the focal point for opposition to Stalin among dissident Communists all over the world, although he was to be more a symbol than an active political force. Having defeated Trotsky and his allies, Stalin next switched policies, abandoning the New Economic Policy in favour of rapid industrialization along with the collectivization of agriculture. The collectivization policy ultimately produced a famine, costing the lives of millions of peasants. The reversal of the New Economic Policy and of Lenin's policy necessarily involved eliminating from the political scene Stalin's former allies, headed by Nikolay Bukharin, who wanted to go slower with industrialization and to cultivate support among the peasants. The protracted conflict, first with Trotsky and his ally G.Y. Zinovyev and then with Bukharin, was reflected in the Comintern and in the world Communist movement, which became increasingly subordinated to Stalin's policy concerns inside the Soviet Union.

Stalin and the Comintern. The regimentation of the Comintern and of the parties represented in it began at the fifth world congress in June 1924, immediately after Lenin's death. The elimination of Trotsky and his supporters within the Soviet party was followed by widespread expulsions of the "left" from the other world parties. The control of the Soviet-dominated Comintern apparatus was increasingly asserted over the tightly disciplined governing bodies of the foreign parties, which in turn ruled over their members with the instrument of the purge. Ideologically, this procedure was carried out at first under the screen of the United Front, which called for cooperation with Social Democrats and other moderate leftists. At the sixth world congress in 1928, however, a further switch in policy was dictated by Stalin's internal conflict: the United Front tactic was abandoned, and the Social Democrats now became enemies along with Fascists. The sixth congress also declared the main duty of the international working-class movement to be the support of the U.S.S.R. by every means. The united front tactic was revived in 1935 at the

seventh (and last) world congress of the Comintern under the name of the Popular Front, calling for united action by Communists and Socialists together against Fascism.

The Nazi–Soviet pact

Comintern policy changed again in August 1939 when the Soviet Union and Germany concluded a 10-year treaty of nonaggression. This had the effect of freeing Hitler to fight a war against Britain and France. Anti-Fascism was now jettisoned, and the Communist parties were required, up to the moment when Germany invaded the Soviet Union on June 22, 1941, to denounce the allied war against Hitler and to recognize Nazism as "the lesser evil" in comparison with Western imperialism. The Soviet alliance with Germany is usually seen as proof that Stalin was primarily concerned with what he considered to be the interests of the Soviet Union. A secret protocol annexed to the treaty assigned the Baltic states (Latvia, Lithuania, and Estonia), about half of Poland, and Bessarabia to the Soviet sphere of influence. The evidence suggests that Stalin considered the deal with Hitler to be based on mutual interests; the German invasion in 1941 took him by surprise. After the defeat of Hitler, Soviet territorial demands were again advanced.

Stalin's method of rule. The Communist parties of the world were also called on to adopt official Soviet justifications for Stalin's internal purges, which involved the extermination of a large proportion of the Soviet party membership, including most of the leading cadres. The subservience of some Communist parties to official assertions made by the Soviet authorities sometimes earned them the reputation of being little more than agents of the Soviet Union inside their own countries, though this did not necessarily diminish their influence or importance in several countries of Europe or in the United States. They found much support among sympathizers with Marxism, who were prepared to overlook Soviet realities in the service of their ideals or of what they considered to be the historical destiny of mankind—in which they saw Stalinism as merely a transitory stage. The Communists and their parties and their contacts provided a valuable recruiting ground for intelligence agents of all kinds prepared to act against their own countries in the interests of Soviet Russia. The effects of Stalin's internal policy on the Communist parties outside the Soviet Union are of vital importance in understanding the attitude adopted by these parties after 1956, when much of Stalin's policy was officially repudiated.

Stalin's method of rule came, by imitation, to be the standard in all other parties. It hinged primarily upon the dominance of his own personality. He ruled over the country in large measure not through the party, as Lenin had, but through personal agents (like Lavrenty Beria, Andrey Vyshinsky, or Georgy Malenkov) and also through the security police (NKVD). The party as an institution declined under Stalin, and between 1934 and 1952 there was only one party congress, in 1939. The general secretaries of the Communist parties abroad imitated Stalin, and strict hierarchical subordination became the way of party life.

GROWTH OF COMMUNISM DURING AND AFTER WORLD WAR II

The wartime prestige of the U.S.S.R.

The undeclared assault by Hitler on the Soviet Union provoked a wave of sympathy for that country among both the open and secret enemies of Hitler in Europe. The Soviet pact with Hitler, and even the manifest blemishes of Stalin's regime, were forgotten: sympathy with the newly emerged force of resistance to the Nazi scourge far outweighed past memories. Many, it is true, expected the immediate defeat of the Soviet Union. As time went on, however, and the Soviet struggle continued with enormous sacrifice of life and with courage and skill that none could help but applaud, admiration for Soviet military achievements grew even among those who had been most critical and apprehensive of the Soviet political role before the war. The Communists of other countries shared in the prestige won by Soviet military prowess. This was particularly the case in occupied France and Italy where the underground Communist parties played a vital role in the resistance movements. In Yugoslavia, too, the Communist partisan movement led by Tito (Josip Broz) outstripped the nationalist guerrillas in effectiveness and won the material support of Britain.

Russian nationalism. The policy pursued by Stalin accentuated the nationalist side of the war and attempted in every way to play down the Communist element. At home, tsarist history and the rituals of the Eastern Orthodox Church were invoked in efforts to raise patriotic sentiments to the highest possible pitch. Abroad, Communist aims and ideals were replaced by anti-Nazi, liberal-democratic slogans. The dissolution of the Comintern in 1943 was in line with this policy. It had long ceased to be necessary as an instrument of Soviet control over the foreign Communist parties, which was carried on through other channels; but the publicizing of its dissolution added force to the growing persuasion abroad that the Soviet Union had left its revolutionary past behind it and was now a great power with traditional nationalist and security aims. Stalin himself emphasized that the dissolution of the Comintern would "put an end to the lies spread by Hitler that the Soviet Union wished to Bolshevize other countries" and that Communist parties "followed foreign directives." Still another factor promoting the influence of Communism during World War II was the enhanced prestige of Stalin himself and the extent to which his personality influenced the allied leaders Winston Churchill and Franklin D. Roosevelt.

Stalin and eastern Europe. His growing military and political prestige in turn influenced Stalin's policy towards his allies and determined the future course of Communism after victory was won in 1945. Two main lines of Soviet policy can be discerned in the wartime conferences at Tehrān, Yalta, and elsewhere: first, a determination by the Soviet Union that friendly political regimes should be established in the countries on Russia's borders, and second, that the Soviet Union's hard-won status as a great power should be fully recognized in the postwar settlements. These demands were not in themselves unreasonable, considering the enormous price that the Soviet people had paid for victory. In pursuing the creation of a solid Soviet-dominated bloc of Communist states in east-central Europe, Stalin was able to take advantage of the presence of a victorious Soviet army in Poland, Bulgaria, Romania, Hungary, and East Germany. The cases of Yugoslavia and Albania were different, but the regimes that emerged in all these countries were broadly similar forms of Communist party domination based on the Soviet model, even though the ways in which the Communists achieved power varied.

The expansion of Soviet influence

Broadly speaking, three phases could be distinguished. In the first phase there was a genuine coalition of Communist and Socialist parties. This lasted until the spring of 1945 in Romania and Bulgaria, until the spring of 1947 in Hungary, and until February 1948 in Czechoslovakia. Yugoslavia, Albania, Poland, and East Germany never knew this phase: the former two started as "monolithic," while the latter two began their postwar history in the second phase, an alleged coalition in which the Socialist parties were nominally independent and had some share in power but in which their leaders and policies were largely determined by the Communists. In the third phase, the "monolithic" phase, the nominally independent Socialist parties were required to fuse with the Communists, political opposition was largely suppressed, and Socialist leaders went into exile or were dealt with by staged treason trials. In Poland, Bulgaria, and Romania the third phase began in the autumn of 1947; in Hungary, in the spring of 1948. In East Germany the third phase was complete by 1949.

In his policy toward the countries which were destined to form the Soviet bloc, Stalin was aided in part by the inability or unwillingness of the Western allied powers to take steps during the first or second phases described above to prevent the beginning of the third phase and in part by the skillful infiltration of local Communists into key positions. The peasant and Socialist parties, which had substantial support in their countries, were attacked in various ways and demolished as independent political bodies.

Yugoslavia was an exception. There the Communists under the leadership of Tito enjoyed a considerable measure

of mass support because of their wartime role as partisan fighters. The People's Democracy they instituted in Yugoslavia was for some years little different in character from that of other Communist-party-dominated states of eastern Europe. An attempt to set up a People's Democracy in Greece failed after three years of civil war, in which the Greek Communists were supported by Yugoslav aid.

The failure in western Europe

In the countries of Europe outside the Soviet bloc, Communist parties proved unable to exploit the prestige that they had acquired during the war. Both in France and in Italy they enjoyed considerable support: in the parliamentary election of 1945 in France the Communists received 26 percent of the vote, and in the general elections to the Constituent Assembly in Italy in June 1946 they received 19 percent. Both parties, however, failed to achieve real national power in the postwar period; their role was confined to fomenting strikes and disorder in the interests of Soviet policy. The detailed story of the Italian and French Communist parties during the period 1945 to 1949 is complex, but, broadly speaking, their attempts at insurrection foundered against the facts of the power of the army and the police and a lack of revolutionary zeal among their worker supporters. On the other hand, their attempts to win power by parliamentary means were frustrated by the distrust that the Socialists felt for them as colleagues in Parliament or in government and by their own evident lack of interest in a viable parliamentary system.

Communism's growth in Asia. Powerful Communist parties emerged after the war in various parts of Asia, in many cases largely as a result of the resistance of the Western powers to growing nationalist movements. Communist-led insurrections, allegedly coordinated by Moscow, broke out in the summer of 1948 in Burma, Malaya, and Indonesia. In Indochina, after the surrender of Japan, the Communists under Ho Chi Minh seized power in the three northern provinces of the country. French colonial policy helped drive the nationalists into the arms of Ho Chi Minh, and by the end of 1946 a guerrilla war had broken out in the country that was to last for nearly three decades before the Communist victory of 1975. In Japan democratic legislation imposed by the United States after its victory permitted the Communists to operate legally. In the succeeding few years they made little progress toward governmental power but won considerable gains in the trade unions and an important measure of influence among university students. In India the Communist Party supported the British war effort after June 1941 and gained ground as a result; it switched to violent insurrection after Indian independence but abandoned this policy in 1950.

Chinese Communism

The most significant factor in the postwar history of Communism in Asia may have been the victory in 1949 of the Chinese Communist Party under the leadership of Mao Tse-tung. China, rather than the Soviet Union, seemed destined to play the leading role in Asian Communism. The victory of the Chinese Communists over Chiang Kai-shek and the Kuomintang, like that of Tito's forces in Yugoslavia, owed little if anything to Soviet aid—save that the Russians had handed over to the Chinese Communists the military stores captured from the Japanese during the very short period when the U.S.S.R. was at war with Japan in 1945. Although the Chinese Communist Party had developed under the aegis of the Comintern and acknowledged the doctrinal authority of Lenin and Stalin, its experience had been very different. Its victory had been preceded by long guerrilla warfare. Mao's rise to power had, moreover, been achieved by ignoring Soviet advice as much as by following it. Stalin showed quite clearly from the outset that he intended to keep China in a position of subordination not unlike that which he had successfully marked out for most of eastern Europe—a status the Chinese Communist leaders were not likely to accept. Culturally, economically, and geographically, China was in a strong position to become the model for Communist revolution in Asia and to wrest the leadership of Asian Communism from the Soviet Union. These and other factors were to produce signs of a possible breach between China and the U.S.S.R. within less than 10 years of the proclamation of the Chinese People's Republic on October 1, 1949.

THE WORLD MOVEMENT UP TO STALIN'S DEATH

The Cold War

The wartime alliance had given rise to some hopes that Soviet-Western amity would continue. Stalin's relentless pursuit of security through the domination of neighbouring countries shattered this hope. At home Stalin returned to his prewar tactics: widespread arrests and deportations occurred in the newly incorporated or reincorporated territories of the Soviet Union; the restriction of cultural life was intensified; the straitjacket was reimposed on the party, on the peasants, and on the industrial workers. There is some evidence to suggest that at the time of his death in March 1953 Stalin was planning a new purge on the scale of the 1936–38 purges.

The struggle with the West. Soviet expansion into eastern Europe led to counteractions by the Western powers that Moscow interpreted as part of a master plan to encircle and subjugate the Soviet Union. These included the Truman Doctrine of containment of Soviet expansion proclaimed in March 1947; the offer in June of that year by United States Secretary of State George Marshall to underwrite the economic recovery of Europe; and the North Atlantic Treaty of April 1949, which established a permanent defense force for western Europe, including in its orbit West Germany. Another factor that affected Soviet policy was the monopoly of the atomic bomb enjoyed by the United States from 1945 until 1949. The Soviet Union rejected the Baruch Plan put forward by the U.S. for the international control of atomic weapons and made every effort to produce its own, succeeding in September 1949. The "Cold War" was on.

The defection of Yugoslavia. In September of 1947 a new international organization, the Communist Information Bureau (Cominform), was established. Unlike the old Third International (Comintern), the Cominform was limited in membership to the Communist parties of the Soviet-dominated countries of east-central Europe and to the French and Italian Communist parties. The aim of the Cominform was to consolidate and expand Communist rule in Europe. Plans for the establishment of Communist rule in Czechoslovakia were discussed, and the French and Italian parties were reproved for their failure to win power in their own countries.

The Cominform did not prove a success. Certainly one of its purposes was to hold Yugoslavia more securely within the Communist fold, and for this reason Belgrade was chosen as the seat of the new organization. But within a few months a quarrel broke out between the Soviet and Yugoslav parties, and when the Cominform held its second meeting in June 1948, it was for the purpose of denouncing the Yugoslav Communist Party and expelling it from the organization. The quarrel with Yugoslavia resulted largely from Tito's refusal to submit to domination by the Soviet Union; there was also some suspicion on the Soviet side, possibly well founded, that the Yugoslav party leader hoped to build up a bloc of Communist states in southeastern Europe that would not be totally dependent on the Soviet Union.

The effect of the Soviet-Yugoslav quarrel, which has never completely healed, was momentous. First, it shattered the doctrine that the Communist movement must be monolithic, since a Communist party had challenged Moscow and survived. Second, Yugoslavia, having broken with the U.S.S.R., was in a position to assume a role of considerable influence in the world, especially toward states formed in formerly colonial territories. The Yugoslavs could speak as Communists who, while opposed to the policy of the imperialist powers, were no mere agents of Soviet policy. This position carried a particularly strong appeal in India, but the impact of the Soviet quarrel with Tito was much wider.

A third effect of the Yugoslav defection was a tightening of the Soviet hold over the remaining members of the Communist bloc. In Soviet-dominated lands "Titoism" became synonymous with treason, much as "Trotskyism" had been in the '30s. Purges and public trials ensued throughout eastern Europe. In some cases, like that of Władysław Gomułka in Poland (who was left alive), or Koci Xoxe in Albania, the charge of sympathy with Yugoslavia may have been true; in others, like those of

László Rajk in Hungary or Traicho Kostov in Bulgaria, the offense may have been only an attempt to resist Soviet domination; in the trial of Rudolf Slánský in Czechoslovakia in 1952, a strong anti-Semitic element played a part. Countries of the Communist bloc were seething with anti-Soviet and nationalist feeling by the time Stalin died. Though Stalin's postwar policy was successful in extending the boundaries of Soviet military and political control well into eastern and central Europe, Communism did not win out in France or in Italy, where its chances had appeared strongest. The policy of expansionism and of intransigence founded on suspicion of the United States led to a kind of consolidation of the West against the Soviet Union. In the Far East the Korean War was probably not a success from the Communist point of view. Korea had been divided after the defeat of Japan: in the northern part a Communist government came to power in elections held in November 1946, and in the south a non-Communist government was established. Each claimed to be the legal government of the whole country. Invasion of the south by the north in June 1950 was condemned by the Security Council of the United Nations as aggression, and the Security Council approved military assistance to South Korea under a unified American command. (The absence of the Soviet representative from the Security Council prevented the U.S.S.R. from vetoing this resolution.) The long war, in which China intervened on the side of North Korea, brought heavy burdens and few, if any, advantages, and the conflict between the major powers that it involved led them in the fears of many to the verge of world war. In June 1951 the Soviet Union proposed discussions for an armistice, to which the Western powers agreed. The negotiations were protracted and did not result in an armistice until after Stalin's death in 1953.

THE BREAKUP OF THE WORLD COMMUNIST MONOLITH

Stalin's heirs

The Khrushchev era. Stalin died on March 5, 1953. For a short time, until the beginning of 1955, power was nominally divided between Georgy Malenkov, the chairman of the Council of Ministers, and Nikita Khrushchev, the first secretary of the Communist Party. Almost from the beginning, Khrushchev was the dominant of the two; his victory over his rival was only a matter of time. Malenkov, it would seem, decided quite early that the Soviet Union could not maintain its hold over the Eastern bloc without substantial economic relaxation. The difficulties that always beset the reform of an oppressive regime were soon illustrated in East Germany. Within a week of the announcement by East German leaders that "aberrations" of the past would be rectified and some of the hardships of life alleviated, there was an uprising in the streets of East Berlin; it spread to other parts of East Germany and was quelled only by the use of Soviet armed forces. The blame for this was laid on Lavrenty Beria (the Soviet security chief, shortly to be deposed and executed) and by implication on Malenkov. The new relaxed policy continued, however, in most of the Soviet-bloc countries. Economic reforms were initiated in Hungary, Czechoslovakia, and Poland, but the system of political rule remained unchanged.

Khrushchev, who by the beginning of 1955 had ousted Malenkov, had a comprehensive vision of how the Eastern bloc should be run. He was determined to find a way out of the straitjacket in which Stalin had confined Soviet life; the outcome was to have momentous consequences for Soviet dependencies abroad, which Khrushchev probably did not at the time foresee. His policy toward the Communist satellite countries may be summarized as one of cooperative integration instead of exploitation, with some degree of economic and political autonomy (under Communist Party leadership). A political and military convention between the European Communist states and the U.S.S.R. (the Warsaw Pact) was signed in May 1955. Khrushchev also sought to redesign the Council for Mutual Economic Assistance, the Communist counterpart of western Europe's Common Market, which Stalin had set up in January 1949: he tried (though with indifferent success) to transform the Council for Mutual Economic Assistance into a device for promoting the division of labour, economic specialization, and technical and financial cooperation among the countries of the bloc.

The crises of 1956. In order to demonstrate that Stalin's policy was a thing of the past, Khrushchev made substantial efforts to effect a reconciliation with Tito and the Yugoslav Communists (against the opposition of some of his colleagues, including Vyacheslav Molotov). An agreement with Yugoslavia in June 1956 recognized that "the conditions of Socialist development are different in different countries" and stated that no Socialist country should impose its views on another. This was a momentous change in policy, since it meant that a country could be described as "Socialist" without being obliged to follow all the practices adopted by the Soviet Union or every Soviet turn in foreign relations.

The reconciliation with Yugoslavia was only one of several important events that made the year 1956 a watershed in the history of Communism. In February, at the 20th congress of the Communist Party, Khrushchev delivered a speech in secret session in which he attacked the period of Stalin's rule in most forthright terms. The speech was not published within the Soviet Union, but its text was widely circulated among Communists both within and outside the Soviet Union and was published by the U.S. State Department. Its effect was enormous. Although the disclosures were neither complete nor entirely new, the fact that Khrushchev had uttered them caused a ferment in the Communist movement that was to prove irreversible. It inaugurated a period of freedom of debate and criticism that had been unknown for a quarter of a century; despite efforts both by Khrushchev and by his successors to keep criticism of the "cult of personality" (the accepted euphemism for Stalin's misdeeds) within bounds, the ferment could not be contained.

Khrushchev's criticism of Stalin

The Hungarian Revolution. In the European Communist countries, Khrushchev's disclosures opened the floodgates of pent-up criticism and resentment against the local Stalin-type leaders. In Hungary, Mátyás Rákosi was ousted as party leader in July 1956 and replaced by Ernő Gerő. But Gerő was unable to contain the rising tide of unrest and discontent, which broke out into active fighting late in October, and appealed for Soviet help. The first phase of the Hungarian Revolution ended in victory for the rebels: Imre Nagy became premier and agreed, in response to popular demands, to establish a multiparty system; on November 1 he declared Hungarian neutrality and appealed to the United Nations. On November 4 the Soviet Union, profiting from the lack of response to Nagy from the Western powers, and from the British and French involvement in action against Egypt, invaded Hungary in force and stopped the revolution. In Poland, where the ferment was also reaching dangerous intensity, the Soviet Union accepted a new party leadership headed by the more moderate Władysław Gomułka. There are believed to have been two reasons for this difference in Soviet policy. One was that in Poland the Communist Party remained in control of the situation. The other was that the invasion and subjugation of Poland would have required a military force several times that required in Hungary.

Polycentrism. Inside the Communist states, the suppression of the Hungarian Revolution had a restraining effect. There was, nevertheless, no return to the Stalinist type of domination and exploitation; a slow evolution followed toward a degree of internal autonomy, even in Hungary. The events of 1956 also had profound effects upon Communists outside the Soviet bloc. There were many resignations after the Hungarian Revolution, and those who remained in the fold began to question both Soviet leadership and the nature of a system that had made the ascendancy of Stalin possible. The most trenchant questioning came from the leader of the Italian Communist Party, Palmiro Togliatti, who concluded that the Soviet pattern could no longer be the model for all other countries and called in June 1956 for decentralization of the Communist movement, a view that became known as "polycentrism." "The whole system becomes polycentric, and . . . we cannot speak of a single guide but rather of a progress which is achieved by following paths which are often different." Although the Italian Communist Party,

or segments of it, were still prepared to support the Soviet Union at times of crisis, at other times it took positions different from those of the Soviet Union.

Opening of the rift with China

The Sino-Soviet dispute. A gathering of Communist parties in Moscow in November 1957, in which China played a leading role, attempted to reassert a common doctrine while recognizing the need for differences in national practice. At Chinese insistence it also retained the Stalinist emphasis on the leadership of the Soviet Union. For a short time relations between the Soviet Union and China were harmonious: after 1955 Khrushchev had put an end to the humiliating terms that Stalin had imposed on China and inaugurated a policy of substantial economic aid.

The differences between China and the Soviet Union, which were to erupt into an open campaign of mutual abuse by 1962, were discernible to most observers by 1959, when the Soviet Union failed to give immediate political backing to Chinese military action against India and when China, at the same time, showed suspicion of Soviet talks with the United States in pursuit of Khrushchev's policy of "peaceful coexistence." In 1960 the differences widened, though they were still unpublicized. The Soviet Union withdrew its technical advisers from China as a preliminary to what was to prove an almost complete severing of economic relations. A facade of agreement was maintained, and at a conference of Communist parties held in Moscow in 1960 a series of resolutions was put forth to show that unity prevailed as ever in the ranks of the world Communist movement. News of serious disagreements, however, soon leaked out, for the increasing number of dissident groups within the several parties had by now rendered the maintenance of secrecy impossible. In the following year, 1961, the Soviet Union began a public polemic against the Chinese viewpoint. This was disguised as an attack on Albania, since 1959 a client of China and increasingly critical of Khrushchev's foreign policy. By 1962 the quarrel had become open and very bitter. It was conducted as a dispute over doctrine, but the practical issue underlying it was a basic rivalry for leadership of the world revolutionary movement.

The Sino-Soviet dispute had three major effects on this movement. It shattered the pretension that Marxism-Leninism offered a single world view, since at least two radically different ways of interpreting Marxism-Leninism were presented to Communists throughout the world, each backed by a Communist party in power with the prestige of a victorious revolution behind it. Second, it seriously impaired, if it did not destroy, the Soviet claim to be the leader of the world revolutionary movement. Since 1960 nearly all Communist parties have split into pro-Soviet and pro-Chinese portions, though outside Asia the Soviet portion has usually retained predominance. In the important parts of Asia, with the possible exception of India, where the party is divided into several warring factions, China has become the predominant influence upon Communist parties. Third, the mere fact of the dispute tended to create greater flexibility for individual parties within the Communist movement as a whole, even in the case of parties that nominally accepted Soviet leadership. The Romanians, for example, were able to follow a nationalistic course by which they successfully resisted Soviet attempts to integrate the Romanian economy into the bloc pattern. The Romanians also took an independent line in their trade relations with other countries, in refusing to participate in the 1968 invasion of Czechoslovakia, and in their policy toward Israel.

After the fall of Khrushchev in October 1964, his successors made efforts to reunite the world movement. They were only moderately successful. Seventy-five parties met in Moscow in June 1969, but of 14 parties in power five did not attend, and Cuba sent only an observer; Asia and Africa, the main areas of Chinese influence, were very poorly represented. Little unity emerged from the conference; in particular, the efforts of the Soviet Union to secure condemnation of China were unsuccessful. The resolution finally adopted was couched in such general terms as scarcely to conceal that the cracks had been merely pasted over. In the course of the 1970s, the hold of the Soviet Communist party over Communist parties outside the bloc seemed for a time to become weaker, with several parties (notably of France, Spain, and Italy) asserting independence from Moscow and the right to criticize Soviet policy. This movement, nicknamed "Eurocommunism," had lost much of its force by the end of the decade, however.

PROBLEMS OF INTERNAL REFORM

The attempt to modify Stalinism

A continuing problem in the history of Communist countries after the death of Stalin was the reform of their overcentralized political and economic structures. The only country that may be said to have achieved success was Yugoslavia, which had since 1948 asserted and maintained its independence from Soviet interference. After initially collectivizing much of its agriculture, Yugoslavia allowed the collective farms to dissolve. It also established Workers' Councils in the factories and publicized them in its foreign propaganda despite Soviet disapproval. The Yugoslav party program of 1958 contained three points in particular that were diametrically opposed to Soviet theory: that Socialism can be achieved without a revolution, that the Communist Party need not have a monopoly of leadership, and that danger of war arises from the existence of two power blocs in the world and not (as the Soviet Union contended) from the aggressive intentions of the United States. In January 1974, a new constitution was adopted that, apart from making changes in the representational system, provided for a collective presidency consisting of one member from each republic and autonomous province. Tito was elected president for life; after his death in 1980 this office rotated among the several members of the collective presidency.

Suppression of reform in Czechoslovakia. The most dramatic failure of an attempt at reform was in Czechoslovakia. The resignation of the old Stalinist party leader Antonín Novotný and his replacement by Alexander Dubček in January 1968 inaugurated a process of liberalization. The reformers hoped to humanize Communist rule by introducing basic civil freedoms, an independent judiciary, and other democratic institutions. The support of leading economists for this program was particularly significant since it indicated that they realized that the already accepted policy of economic decentralization (which included giving a measure of initiative to individual enterprises) would fail unless accompanied by political changes.

Invasion of Czechoslovakia

While the Czechoslovak Communists had repeatedly declared their intention to remain within the existing system, Moscow, possibly fearing that the developments they had set under way would ultimately endanger the stability of eastern Europe, endeavoured to induce the Czechoslovak party leaders to abandon their course. The Soviet effort failed, possibly because there were no Czechoslovak Communist leaders prepared, with Soviet help, to oust Dubček. Finally a group of Warsaw Pact forces—predominantly Soviet, but with token contributions from the other Warsaw Pact members except Romania—invaded Czechoslovakia on the night of August 20–21, 1968, effectively killing the momentum of the reform movement in Czechoslovakia. A Soviet-controlled security service was installed, and the Dubček leadership was gradually forced out of top posts and eventually expelled from the party. Although the repression was thorough, there was no mass terror.

The Soviet invasion of Czechoslovakia came as a greater shock to many Communists than the invasion of Hungary because it was directed against Communist leaders who strongly asserted their loyalty to Moscow. The motives that prompted Soviet action were probably two: one was the fear that the Soviet defense area created by Stalin after World War II might be endangered if the Dubček regime were allowed to continue; the other was the fear that the entrenched and conservative Communist parties in other European Communist countries, and in the Soviet Union itself, might not be equal to the challenge posed by a reformed Communism in Czechoslovakia.

Khrushchev's reforms. This concern that the power of the Communist party might be diminished may also have acted as a brake on internal reform. The reforms carried out by Khrushchev between 1953 and 1964 had been extensive. The arbitrary powers of the security police were

brought under control; there were widespread reviews and rehabilitations (often posthumously) of the sentences of those sent to labour camps under Stalin; and reforms (in 1958) removed the worst anomalies of Soviet criminal law and procedure. The stringent controls over the lives of workers and farmers were relaxed. Discussion and debate were tolerated among writers and intellectuals to a degree that would have been inconceivable under Stalin. The whole system of agricultural management was considerably relaxed, and a system of incentives for collective farmers was introduced. The limit of reform, as Khrushchev saw it, was the point at which any threat appeared to the party's control over all aspects of life. Under his successor, Leonid Brezhnev, the brake on reform was applied more heavily. Criticism of Stalin decreased. Freedom of opinion was considerably restricted by the introduction of penal provisions against "slandering" the Soviet system: for the first time since Stalin's death, there were trials of writers, and the courts ceased to show any inclination to assert their independence as they had under Khrushchev. The numbers of political prisoners steadily increased, although the Brezhnev regime could not be compared to Stalin's. A movement toward economic reform had started under Khrushchev, aiming at some decentralization of economic control through greater freedom for enterprises to plan their own operations and through more influence for market forces. This was continued and officially encouraged after 1964 by Prime Minister Aleksey Kosygin, but it made little headway and was abandoned. The period of the 1970s was one of economic stagnation and conservatism coupled with expanded military power.

Brezhnev's policy toward reform

COMMUNIST DOCTRINE AFTER STALIN

The errors of "revisionism" and "dogmatism." The most far-reaching innovation in Communist doctrine during the period 1953–70 was the Chinese interpretation of Marxism-Leninism known as Maoism. In the Soviet sphere several profound changes in doctrine took place after the death of Stalin. One change was the rise of ideological dispute for the first time since the early 1920s. The Yugoslav ideas were denounced as "revisionism," a term that harked back to the turn of the century when it had been used to characterize the views of Eduard Bernstein, who had argued that Socialism could be achieved without a revolution. After 1957 the terms "revisionism" and "dogmatism" became an integral part of Communist discourse. They were applied in a variety of meanings. By the Chinese, "revisionism" was used to mean, in effect, Khrushchevism—*i.e.*, the policies that Khrushchev had introduced in both domestic and international relations and that the Chinese opposed. On the Soviet side, "revisionism" became a catchphrase to designate any political reform that appeared to endanger the dominance of the Communist Party. As defined at the Moscow conference of 1957 (with Chinese approval then), it was applied to all reform movements within the Communist system that denied "the historical necessity of the proletarian revolution," or the "Leninist principles for the construction of the party." The term "dogmatism," in Soviet usage, meant a doctrinal conservatism that ignores changing realities, a clinging to received ideas in a way "calculated to alienate the party from the masses." In practice the Soviets sought a course between revisionism and dogmatism.

Different roads to Socialism. Important new elements in Soviet doctrine were set out in the party program adopted by the 22nd congress in October 1961 (which were, to some extent, embodied in the declarations of the Moscow conferences of 1957 and 1960). First, there was the concession that there are different roads to Socialism. This may have been no more than a practical recognition of the fact that since the breach with Yugoslavia and the death of Stalin it had no longer been possible for the Soviet Union to impose its own pattern on all Communist states. The invasion of Hungary in 1956, of Czechoslovakia in 1968, and of Afghanistan in 1979 were not, according to Moscow, inconsistent with this doctrine, since in each case the Soviet Union acted out of a duty to assist a fraternal Socialist state in putting down a counterrevolution. In the case of Czechoslovakia, which had not asked for such assistance, a new tenet was added by Brezhnev in November 1968. Known as the "Brezhnev Doctrine," it contended that attempts by "internal and external" forces hostile to Socialism to restore capitalism in a Socialist country were a matter of concern to the whole Socialist community. This tenet was used to justify the action of the Warsaw Pact forces in 1968 and of the Soviet forces in 1979.

The second change in Soviet doctrine was the view that war between the capitalist and Socialist powers was no longer inevitable, as had always been asserted by both Lenin and Stalin. This was a practical recognition of the fact that a war waged with nuclear weapons would be more likely to lead to mutual annihilation than to victory. Khrushchev emphasized the possibility of "peaceful coexistence" between different social systems and the achievement of Socialism by peaceful means. In the 1970s, peaceful coexistence became known as "détente." This doctrine raised hopes of real peace between Communist and non-Communist states, but the Soviet leaders made it clear that détente would not preclude either political warfare against the West or military support for wars of liberation. The Soviet invasion of Afghanistan in 1979 left détente seriously impaired.

"Peaceful coexistence"

The third doctrinal change after 1953 was also dictated by practical reality. The Comintern had rigidly applied concepts drawn from Western history to revolutions in Africa and Asia: industrialization, the emergence of a proletariat, and a Socialist revolution carried out under the leadership of a Communist party. This Marxist analysis proved to be totally unrealistic in the case of underdeveloped countries in which the predominant force was nationalism. This was increasingly recognized, after 1956, in Soviet doctrine that declared the proper revolutionary aim in the developing countries to be "national democracy." In Khrushchev's words this meant accepting a "noncapitalist path of development," which would be in the interests "not only of one class but of the broad strata of the people."

In the late 20th century the Soviet leadership faced two main problems: a decline in the rate of economic growth, to which the party had tied its promises of an improved standard of living, and a ferment of criticism among an intellectual minority, which included an influential component of leading scientists. Two alternatives seemed the most likely: either a return to more repressive measures or a reform of the Soviet system. (L.B.S.)

The collapse of Soviet Communism. Following the death of Brezhnev in 1982, a new generation of less dogmatic party technocrats chose reform. Led by Mikhail S. Gorbachev, who became general secretary in 1985 and president in 1988, Soviet leaders spoke of basic structural reform (*perestroika*) and more openness (*glasnost*) in Soviet society and in foreign policy. In what amounted to a fourth doctrinal change, Soviet leaders declared that Communist revolution was no longer the mission of the Soviet Union, nor would the country continue to serve as the ideological model for world Communism. Underscoring this doctrinal reversal, the Communist Party officially gave up its monopoly of power at the 28th party conference in 1990. The more relaxed attitude in Soviet society subsequently encouraged Soviet-bloc countries in eastern Europe and Africa to develop a more independent stance, and in fact many of them cast out their Communist leaders altogether. The dramatic reversal of fortune experienced by the Polish labour movement Solidarity is a prime example. Although it was outlawed by the Communist authorities in 1980, only 10 years later its leader, Lech Wałęsa, became president of Poland.

Perestroika and *glasnost*

Fearing the disintegration of the Soviet Union, in August 1991 a group of hard-line Communist officials detained Gorbachev and attempted to take control of the government. The coup failed after only three days, further encouraging the constituent republics to secede and dealing a deathblow to the already weakened Communist Party. On December 8, 1991, the leaders of Russia, Ukraine, and Bellorussia (now Belarus) declared that the Soviet Union had ceased to exist. Gorbachev resigned as Soviet president on December 25, and all Soviet institutions ceased to function at the end of the year.

Reforms in China. Like the Soviet Union, China also un-

derwent fundamental shifts in policy in the 20th century. Following the economic failures of the Cultural Revolution (1966–76), it adopted a modernization plan designed to attract foreign investment; to improve agriculture, industry, science and technology, and defense; to allow greater individual freedom of choice; and to reduce the influence of political dogmatism in nonpolitical spheres of life. The economy, especially in South China, grew at a record pace from the late 1980s as the government introduced extensive free-market reforms, which were expanded further at a Communist Party plenum in November 1993. In March 1999 the People's Congress adopted two constitutional amendments, one affirming that private enterprise is "an important component of the socialist economy," the other stating that the country "should implement the principle of rule by law." (Ed.)

Anarchism

Anarchism describes a cluster of doctrines and attitudes united in the belief that government is both harmful and unnecessary. Derived from a Greek root signifying "without a rule," anarchism, anarchist, and anarchy are used to express both approval and disapproval. In early contexts all these terms were used pejoratively: during the English Civil Wars of the 17th century the opponents of the radical Levellers referred to them as "Switzerising anarchists," and during the French Revolution the Girondin leader Jacques-Pierre Brissot accused his most extreme rivals, the Enragés, of being the advocates of "anarchy."

> Laws that are not carried into effect, authorities without force and despised, crime unpunished, property attacked, the safety of the individual violated, the morality of the people corrupted, no constitution, no government, no justice, these are the features of anarchy.

These words uttered by the leader of the French Revolutionary moderates in 1793 could serve as a model for the denunciations delivered by all opponents of the anarchists. The latter, for their part, would admit many of Brissot's points. They deny man-made laws, regard property as a means of tyranny, and believe that crime is merely the product of a society based on property and authority. But they would argue that their denial of constitutions and governments leads not to "no justice" but to the real justice inherent in the free development of man's sociality, his natural inclination, when unfettered by laws, to live according to the principles and practice of mutual aid.

ANARCHIST THINKERS

Pierre-Joseph Proudhon

The first person to willingly call himself an Anarchist was the French political writer and pioneer Socialist Pierre-Joseph Proudhon. In 1840, writing his controversial study of the economic bases of society, *Qu'est ce que la propriété?* (*What Is Property?*), Proudhon set out to shock his readers into attention by declaring: "I am an anarchist!" He went on to explain that in his view the real laws of society have nothing to do with authority but stem from the nature of society itself; accordingly, he foresaw the eventual dissolution of authority and the emergence of a natural social order.

> As man seeks justice in equality, so society seeks order in anarchy. Anarchy—the absence of a sovereign—such is the form of government to which we are every day approximating.

The essential elements of the philosophy to which Proudhon in 1840 gave the name of Anarchism had already been developed by various earlier thinkers. There is a tradition of the rejection of political authority going back to classical antiquity, to the Stoics and the Cynics, and recurring throughout Christian history in dissenting sects such as the medieval Catharists and certain factions of Anabaptists during the Reformation. With such groups—often mistakenly claimed as ancestors by modern anarchist writers—the rejection of political government is merely one aspect of a retreat from the material world into a realm of spiritual grace; it becomes part of the search for individual salvation and as such is hardly compatible with the sociopolitical doctrine of Anarchism. That doctrine in all its forms consists of (1) a fundamental criticism of the existing power-based order of society, (2) a vision of an alternative libertarian society based on cooperation as opposed to coercion, and (3) a method of proceeding from one order to the other.

English anarchist thought. The first sketch of an anarchist commonwealth in this sense was developed in the years immediately following the English Civil Wars by Gerrard Winstanley, a dissenting Christian who identified God with reason and founded the minute Digger movement. In his pamphlet of 1649, *Truth Lifting Up Its Head Above Scandals,* Winstanley laid down what later became basic principles among the anarchists: that power corrupts; that property is incompatible with freedom; that authority and property are between them the begetters of crime; and that only in a society without rulers, where work and its products are shared, can men be free and happy, acting not according to laws imposed from above but according to their consciences. Winstanley was not only the pioneer of anarchist theory but also the forerunner of anarchist activism. He held that only by their own deeds can the people bring an end to social injustice, and in 1649, calling upon the people "to manure and work upon the common lands," he led a band of his followers in occupying a hillside in southern England, where they set about cultivation, established free communism among themselves, and offered passive resistance to the hostile landlords.

The Digger experiment was destroyed by local opposition, and Winstanley himself vanished into such obscurity that the place and date of his death are unknown. But the principles he defended lingered on in the traditions of English Protestant sects and reached their ultimate flowering in the masterpiece of a former dissenting minister, William Godwin, who in 1793 published his *Political Justice*—of which the political economist Sir Alexander Gray said that in it "Godwin sums up, as no one else does, the sum and substance of anarchism, and thus embodies a whole tradition." This is essentially true, since Godwin not only presents the classic anarchist argument that authority is against nature and that social evils exist because men are not free to act according to reason, but he also sketches out a decentralized society in which the small autonomous community (the parish) is the essential unit. In his community, democratic political procedures are dispensed with as far as possible, because even the rule of the majority is a form of tyranny, and such procedures as voting dilute the responsibility of the individual. Godwin also condemns "accumulated property" as a source of power over others and envisages a loose economic system in which men will give and take according to their needs. Godwin was a prophet of technological progress; he believed that industrial development would eventually reduce the necessary working time to half an hour a day, provided men lived simply, and that this would facilitate the transition to a society without authority.

Godwin's classic argument

Godwin enjoyed great celebrity in the 1790s and influenced such varied writers as Percy Bysshe Shelley (whose *Queen Mab* and *Prometheus Unbound* are virtually anarchist poems), William Wordsworth, William Hazlitt, and Robert Owen, but he was almost forgotten by the time of his death in 1836. Though his ideas were to have, through Owen, a subterranean influence on the British labour movement, it is only recently that professed Anarchists have recognized his affinities with them.

French anarchist thought. The theoretical foundations of the Continental anarchist movement were laid by Pierre-Joseph Proudhon. A brewer's son of peasant stock from the Franche-Comté, he started life (like many later anarchists) as a printer, but by the revolutionary year of 1848 he had already become a polemicist and a radical journalist with two books to his credit, *Qu'est ce que la propriété?* and *Système des contradictions économiques* (*System of Economic Contradictions*). These works established him among the leading theoreticians of Socialism, a term that in the early 19th century embraced a wide spectrum of attitudes. In Paris during the 1840s, Proudhon associated with Karl Marx and the Russian Mikhail Bakunin, and, out of the experiences of the Revolutions of 1848 (when he served in the Constituent Assembly and voted against the constitution "because it is a constitution"), he developed the libertarian theories that he discussed in later works such

as *Du principe fédératif* (*The Federal Principle*) and *De la capacité politique des classes ouvrières* (*The Political Capability of the Working Classes*).

The work of Proudhon

Proudhon was a complex and voluminous writer who remained obstinately independent, refusing to consider himself the founder of either a system or a party. Yet he was justly regarded by Bakunin, Peter Kropotkin, and other leaders of organized Anarchism as their true ancestor, for he had adumbrated their philosophy.

Mutualism, federalism, and direct action were the essential doctrines Proudhon taught. By mutualism he meant the organization of society on an egalitarian basis. He declared that "property is theft," but this did not mean that he advocated communism. He attacked the use of property as a means of exploiting the labour of others, but he regarded "possession"—the right of a worker or group of workers to control the land or tools necessary for production—as an essential bulwark of liberty. He therefore envisaged a society formed of independent peasants and artisans, with factories and utilities run by associations of workers, all united by a system of mutual credit founded on people's banks. In place of the centralized state—the enemy of all Anarchists—Proudhon suggested a federal system of autonomous local communities and industrial associations, bound by contract and mutual interest rather than by laws, with arbitration replacing courts of justice, workers' management replacing bureaucracy, and integrated education replacing academic education. Out of such a network would emerge a natural social unity compared with which the existing order would appear as "nothing but chaos, serving as a basis for endless tyranny."

Proudhon remained all his life an independent polemicist, but in his final, posthumously published work, *De la capacité politique des classes ouvrières,* with its insistence that the liberation of the workers must be the task of the workers themselves organized in industrial associations, he laid the intellectual foundations of a movement that would reject democratic and parliamentary politics in favour of various forms of direct action. Unlike their master, Proudhon's working-class followers of the 1860s did not accept the title Anarchist (though in 1850 an independent revolutionary, Anselme Bellegarrigue, had founded a short-lived magazine entitled *L'Anarchie*); they preferred to call themselves Mutualists, after a working-class secret society bearing the same name to which Proudhon had belonged in Lyons during the 1830s. In 1864, shortly before Proudhon's death, a group of them joined with British trade unionists and European Socialists exiled in London to found the International Workingmen's Association (the First International). The Mutualists became the first opposition within the International to Karl Marx and his followers, who advocated political action and the seizure of the state in order to create a proletarian dictatorship. Marx's most formidable opponents, however, were not the Mutualists but the followers of Mikhail Bakunin, a Russian nobleman turned revolutionary who entered the International in 1868 after a long career as a political conspirator.

The work of Bakunin

Russian anarchist thought. Bakunin had begun as a supporter of nationalist revolutionary movements in various Slav countries. In the 1840s he had come under the influence of Proudhon, and by the 1860s, when he entered the International, he had not only founded his own proto-Anarchist organization, the Social Democratic Alliance, with a considerable following in Italy, Spain, Switzerland, and the Rhône Valley, but had modified the Proudhonian teachings into the doctrine that became known as collectivism. Bakunin accepted Proudhon's federalism and his insistence on the need for working-class direct action, but he argued that the modified property rights Proudhon allowed were impractical and instead suggested that the means of production should be owned collectively, though he still held that each worker should be remunerated only according to the amount of work he actually performed. The second important difference between Bakunin and Proudhon lay in their concepts of revolutionary method. Proudhon believed it was possible to create within existing society the mutualist associations that could replace it; he therefore opposed violent revolutionary action. Bakunin, declaring that "the passion for destruction is also a creative urge," refused to accept the piecemeal approach; a violent revolution, sweeping away all existing institutions, was in his view the necessary prelude to the construction of a free and peaceful society.

Though the individualism and nonviolence implicit in Proudhon's vision have survived in the peripheral currents of the anarchist tradition, it was Bakunin's stress on collectivism and violent revolutionary action that dominated the mainstream from the days of the First International down to the destruction of anarchism as a mass movement at the end of the Spanish Civil War in 1939.

The First International was itself destroyed by the conflict between Marx and Bakunin, a conflict rooted as much in the contradictory personalities of the two leaders as in their rival doctrines—revolution by a disciplined party versus revolution by the spontaneous insurgence of the working class. When the international finally broke apart at the Hague congress in 1872, Bakunin's followers were left in control of the working-class movements in the Latin countries—Spain, Italy, southern France, and French-speaking Switzerland—and these were to remain the principal bases of Anarchism in Europe. In 1873 the Bakuninists set up their own International, which lasted as an active body until 1877; during this period its members finally accepted the title of Anarchist rather than collectivist.

Bakunin died in 1876. His ideas had been developed in action rather than in writing, for he was the hero of many barricades, prisons, and meetings. His successor as the ideological leader, Peter Kropotkin (who renounced the title of prince when he became a revolutionary in 1876), is more celebrated for his writing than for his actions, though in his early years he led an eventful career as a revolutionary militant, which he described in a fine autobiography, *Memoirs of a Revolutionist* (1899). Under the influence of the French geographer Elisée Reclus (a former disciple of the Utopian Socialist Charles Fourier), Kropotkin developed the variant of anarchist theory known as anarchist communism. Kropotkin and his followers went beyond Bakunin's collectivism, since they argued not only that the means of production should be owned cooperatively but also that there should be complete communism in terms of distribution; this revived the scheme Sir Thomas More had sketched out in his 16th-century *Utopia* of common storehouses from which everyone should be allowed to take whatever he wished on the basis of "From each according to his means, to each according to his needs." In *La Conquête du pain* (*The Conquest of Bread,* 1892) Kropotkin sketched the vision of a revolutionary society organized as a federation of free communist groups. He reinforced the vision by writing *Mutual Aid: A Factor in Evolution* (1902), in which he endeavoured to prove by means of biological and sociological evidence that cooperation is more natural and usual among both animals and men than competition. In his *Fields, Factories and Workshops* (1899) he put forward ideas on the decentralization of industry appropriate to a nongovernmental society.

The work of Kropotkin

ANARCHISM AS A MOVEMENT

Kropotkin's writings completed the theoretical vision of the Anarchist future, and little new has been added since his time. But this work was of less immediate importance than the emergence among the Italian anarchists of the theory of "propaganda of the deed." In 1876 Errico Malatesta expressed the belief of the Italian Anarchists that "the *insurrectionary deed,* destined to affirm socialist principles by acts, is the most efficacious means of propaganda." The first acts were rural insurrections, intended to arouse the illiterate masses of the Italian countryside. After the insurrections failed, Anarchist activism tended to take the form of individual deeds of protest by terrorists, who would attempt to kill ruling figures in the hope of demonstrating the vulnerability of the structure of authority and inspiring the masses by their self-sacrifice. In this way, between 1890 and 1901, a series of symbolic murders was enacted; the victims included King Umberto I of Italy, the empress Elizabeth of Austria, President Carnot of France, President McKinley of the United States, and Antonio Cánovas del Castillo, the prime minister of Spain. This brief but dra-

Anarchism and terrorism

matic series of terrorist acts established the image of the Anarchist as a mindless destroyer; after 1901, however, the Anarchists continued to practice widespread terrorism only in such countries as Spain and Russia, where the general political atmosphere was conducive to violence.

During the 1890s, especially in France, Anarchism was adopted as a philosophy by avant-garde painters and writers. Gustave Courbet had already been a disciple of Proudhon; among those who in the 1890s accepted an Anarchist philosophy were Camille Pissarro, Georges Seurat, Paul Signac, Paul Adam, Octave Mirbeau, Laurent Tailhade, and, at least as a strong sympathizer, Stéphane Mallarmé. At the same time in England, Oscar Wilde declared himself an Anarchist and, under Kropotkin's inspiration, wrote his libertarian essay, "The Soul of Man Under Socialism" (1891).

The artists were attracted by the individualist spirit of Anarchism. By the mid-1890s, however, the more militant Anarchists in France began to realize that an excess of individualism had tended to detach them from the workers they sought to liberate. Anarchists, indeed, have always found it difficult to reconcile the claims of general human solidarity with the demands—equally insistent—of the individual who desires freedom. Some Anarchist thinkers, such as the German Max Stirner, who published *Der Einzige und sein Eigentum* (*The Ego and His Own*) in 1845, have refused to recognize any limitation on the individual's right to do as he will or any obligation to act socially; and even those who accepted Kropotkin's socially oriented doctrines of Anarchist communism have in practice been reluctant to create forms of organization that threatened their freedom of action or seemed likely to harden into institutions.

In consequence, although a number of international Anarchist congresses were held (the most celebrated being those of London in 1881 and of Amsterdam in 1907), no effective worldwide organization was created, even though by the end of the century the Anarchist movement had spread to all continents and was united by informal links of correspondence and friendship between the leading figures. National federations were weak even in countries where there were many Anarchists, such as France and Italy, and the typical unit of organization was the small group dedicated to propaganda by deed or word. Such groups engaged in a wide variety of activities; in the 1890s many of them concentrated on setting up experimental schools and communities that attempted to live out Anarchist principles.

Revolutionary Syndicalism. In France, where individualist trends had been most pronounced and public reaction to terrorist acts had imperilled the very existence of the movement, an effort was made to acquire a mass following. The Anarchists infiltrated the trade unions. They were particularly active in the *bourses du travail* ("labour exchanges"), local groupings of unions, originally set up to find work for their members, that appealed to the Anarchist ideal of decentralization. In 1892 a national confederation of *bourses du travail* was formed, and by 1895 the Anarchists, led by Fernand Pelloutier, Émile Pouget, and Paul Delesalle, had gained effective control and were developing the theory and practice of working-class activism that became known as Anarcho-Syndicalism, or Revolutionary Syndicalism.

Anarchism and the workers

The Anarcho-Syndicalists argued that the traditional function of trade unions—to struggle for better wages and working conditions—was not enough. The unions should become militant organizations dedicated to the destruction of capitalism and the state. They should aim to take over factories and utilities, which would then be operated by the workers. In this way the union or syndicate would have a double function—as an organ of struggle under the present dispensation and as an organ of administration after the revolution. To sustain militancy, an atmosphere of incessant conflict should be induced, and the culmination of this strategy should be the general strike. Many of the Syndicalists believed that such a massive act of noncooperation would bring about what they called "the revolution of folded arms," resulting in the collapse of the state and the capitalist system. But, although partial general strikes, with limited objectives, were undertaken in France and elsewhere with varying success, the millennial general strike aimed at overthrowing the social order in a single blow was never attempted. The Anarcho-Syndicalists acquired great prestige among the workers of France and, later, of Spain and Italy, because of their generally tough-minded attitude at a time when working conditions were bad and employers tended to respond brutally to union activity. After the great French trade-union organization, the Confédération Générale du Travail (CGT), was founded in 1902, their militancy enabled the Anarchists to retain control of the organization until 1908 and to wield considerable influence on its activities until after World War I.

Like Anarchism, Revolutionary Syndicalism proved attractive to certain intellectuals, notably Georges Sorel, whose *Réflexions sur la violence* (1908; Eng. trans., *Reflections on Violence*, 1914 and 1950) was the most important literary work to emerge from the movement. He argued the importance of the general strike as a social myth. The more purist Anarchist theoreticians were disturbed by the monolithic character of Syndicalist organizations, which they feared might create powerful interest structures in a revolutionary society. At the International Anarchist Congress at Amsterdam in 1907, a crucial debate on this issue took place between the young Revolutionary Syndicalist Pierre Monatte and the veteran Anarchist Errico Malatesta. It defined a division of outlook that still lingers in what remains of the historic Anarchist movement, which has always included individualist attitudes too extreme to admit any kind of large-scale organization.

Revolutionary Syndicalism did transform Anarchism, for a time at least, from a tiny minority current into a movement with considerable mass support, even though most members of Syndicalist unions were sympathizers and fellow travellers rather than committed Anarchists. In 1922 the Syndicalists set up their own International with its headquarters in Berlin, taking the historic name of the International Workingmen's Association; it still survives, with headquarters in Stockholm. When it was established the organizations that formed it could still boast considerable followings. The Unione Sindicale Italiana had 500,000 members; the Federación Obrera Regional Argentina, 200,000 members; the Portuguese Confederação General de Trabalho, 150,000 members; and the German Freie Arbeiter Union, 120,000 members. There were smaller organizations in Chile, Uruguay, Denmark, Norway, Holland, Mexico, and Sweden. In Britain the influence of Syndicalism was shown most clearly in the Guild Socialist movement that flourished briefly in the early years of the present century. In the United States, Revolutionary Syndicalist ideas were manifested in the Industrial Workers of the World (IWW), which in the years immediately before and after World War I played a vital part in organizing American miners, loggers, and unskilled workers; but only a small minority of the IWW militants were ever avowed Anarchists.

Syndicalist organizations

Anarchism in Spain. The reconciliation of Anarchism and Syndicalism was most complete and most successful in Spain; for a long period the Anarchist movement in that country remained the most numerous and the most powerful in the world. The first known Spanish Anarchist, Ramón de la Sagra, a disciple of Proudhon, founded the world's first Anarchist journal, *El Porvenir*, in La Coruña in 1845; it was quickly suppressed. Mutualist ideas were later publicized by Pi y Margall, a federalist leader and the translator of many Proudhon books; during the Spanish revolution of 1873, Pi y Margall attempted to establish a decentralist, or "cantonalist," political system on Proudhonian lines. In the end, however, the influence of Bakunin was stronger. In 1868 his Italian disciple, Giuseppe Fanelli, visited Barcelona and Madrid, where he established branches of the International. By 1870 they had 40,000 members, and in 1873 the movement numbered about 60,000, organized mainly in working men's associations. In 1874 the Anarchist movement in Spain was forced underground, a phenomenon recurring often in subsequent decades. It flourished, nevertheless, and Anarchism became the favoured type of radicalism among two

very different groups, the factory workers of Barcelona and other Catalan towns and the impoverished peasants who toiled on the absentee-owned estates of Andalusia.

As in France and Italy, the movement in Spain during the 1880s and 1890s was inclined toward insurrection (in Andalusia) and terrorism (in Catalonia). It retained its strength in working-class organizations because the courageous and even ruthless Anarchist militants were often the only leaders who would stand up against the army and the employers, who hired squads of gunmen to engage in guerrilla warfare with the Anarchists in the streets of Barcelona. The workers of Barcelona were finally inspired by the success of the French CGT to set up a Syndicalist organization, Solidaridad Obrera (Workers' Solidarity). Established in 1907, Solidaridad Obrera quickly spread throughout Catalonia, and in 1909, when the Spanish army tried to conscript Catalan reservists to fight against the Riffs in Morocco, it called a general strike. "La Semana Tragica," "the Tragic Week" of largely spontaneous violence that followed (with hundreds dead and 50 churches and monasteries destroyed), ended in violent repression. Tortures of Anarchists in the fortress of Montjuich and the execution of the internationally celebrated advocate of free education Francisco Ferrer led to worldwide protests and the resignation of the conservative government in Madrid. These events also resulted in a congress of Spanish trade unionists at Seville in 1910, which founded the Confederación Nacional del Trabajo (CNT).

The CNT, which included the majority of organized Spanish workers, was dominated throughout its existence by the Anarchist militants; these in 1927 founded their own activist organization, the Federación Anarquista Iberica (FAI). While there was recurrent conflict within the CNT between moderates and FAI activists, the atmosphere of violence and urgency in which radical activities were carried on in Spain ensured that the more extreme leaders, such as Garcia Oliver and Buenaventura Durutti, tended to wield the decisive influence. The CNT was a model of Anarchist decentralism and antibureaucratism: its basic organizations were not national unions but *sindicatos únicos,* which brought together the workers of all trades and crafts in a certain locality; the national committee was elected each year from a different locality to ensure that no individual served more than one term; and all delegates were subject to immediate recall by the members. This enormous organization, which claimed 700,000 members in 1919, 1,600,000 in 1936, and more than 2,000,000 during the civil war, employed only one paid secretary. Its day-to-day work was carried on in their spare time by workers chosen by their fellows. This meant that the Spanish Anarchist movement was not dominated by the *déclassé* intellectuals and self-taught printers and shoemakers who were so influential in other countries. One consequence was that the Spanish movement contributed nothing original to the ideological literature of Anarchism; like Bakunin, whom they favoured most among the classic libertarian thinkers, the Spanish Anarchists developed their attitudes in action rather than in writing.

Moderates and activists

The CNT and the FAI, which remained clandestine organizations under the dictatorship of Primo de Rivera, emerged into the open with the abdication of King Alfonso XIII in 1931. Their antipolitical philosophy led them to reject the republic as much as the monarchy it had replaced, and between 1931 and the military rebellion led by Francisco Franco in 1936 there were several unsuccessful Anarchist risings. In 1936 the Anarchists, who over the decades had become expert urban guerrillas, were mainly responsible for the defeat of the rebel generals in both Barcelona and Valencia, as well as in country areas of Catalonia and Aragon; and for many early months of the Civil War they were in virtual control of eastern Spain, where they regarded the crisis as an opportunity to carry through the social revolution of which they had long dreamed. Factories and railways in Catalonia were taken over by workers' committees, and in hundreds of villages in Catalonia, Levante, and Andalusia the peasants seized the land and established libertarian communes like those described by Kropotkin in *La Conquête du pain.* The internal use of money was abolished, the land was tilled in common, and village products were sold or exchanged on behalf of the community in general, with each family receiving an equitable share of food and other necessities. An idealistic Spartan fervour characterized these communities, which often consisted of illiterate labourers; intoxicants, tobacco, and sometimes even coffee were renounced; and millenarian enthusiasm took the place of religion, as it has often done in Spain. The reports of critical observers suggest that at least some of these communes were efficiently run and more productive agriculturally than the villages had been previously.

Anarchists in the Spanish Civil War

The Spanish Anarchists failed during the Civil War largely because, expert though they were in spontaneous street fighting, they did not have the discipline necessary to carry on sustained warfare; the columns they sent to various fronts were unsuccessful in comparison with the Communist-led International Brigades. In December 1936 four leading Anarchists took posts in the Cabinet of Francisco Largo Caballero, radically compromising their antigovernmental principles. They were unable to halt the trend toward left-wing totalitarianism encouraged by their enemies the Communists, who were numerically far fewer but politically more influential. In May 1937 bitter fighting broke out in Barcelona between Communists and Anarchists. The CNT held its own on this occasion, but its influence quickly waned. The collectivized factories were taken over by the central government, and many agricultural communes were destroyed by Franco's advance into Andalusia and by the hostile action of General Lister's Communist army in Aragon. In January 1939 the Spanish Anarchists were so demoralized by the compromises of the Civil War that they were unable to mount a resistance when Franco's forces marched into Barcelona; the CNT and FAI became phantom organizations in exile.

Decline of the Anarchist movement. By this time the movement outside Spain had been either destroyed or greatly diminished as a result of two developments: the Russian Revolution and the rise of right-wing totalitarian regimes. Though the most famous Anarchist leaders, Bakunin and Kropotkin, had been Russian, the Anarchist movement had never been strong in Russia, partly because the more numerous Socialist Revolutionary Party (the Narodniki) had adopted Bakuninist ideas while remaining essentially a constitutional party. After the 1917 revolution the small anarchist groups that emerged in Petrograd (now Leningrad) and Moscow were powerless against the Bolsheviks, and Kropotkin, who returned from exile, found himself without influence. Only in the south did N.I. Makhno, a peasant Anarchist, raise an insurrectionary army that by brilliant guerrilla tactics held a large part of the Ukraine from both the Red and the White armies; but the social experiments developed under Makhno's protection were rudimentary, and when he was driven into exile in 1921 the Anarchist movement became extinct in Russia.

The competition of Communism

In other countries, the prestige of the Russian revolution enabled the new Communist parties to win much of the support formerly given to the Anarchists, particularly in France, where the CGT passed permanently into Communist control. The large Italian Anarchist movement was destroyed by Benito Mussolini's Fascist government in the 1920s, and the small German Anarchist movement was smashed by the Nazis in the 1930s. In Japan, Anarchism had emerged during the Russo-Japanese war of 1904–05, when the Socialist leader Shusui Kotoku became converted by reading Kropotkin in prison; Kotoku and other Anarchists were executed in 1911 for their involvement in a plot against the Emperor, but, after World War I, new Anarchist organizations appeared, including a Black Federation and a Syndicalist federation. After the Japanese invasion of Manchuria in 1931, the imperial government began to suppress all left-wing groups, and the Anarchist movement was finally destroyed in 1935 after a secret society, the Anarchist Communist Party, had been accused of plotting armed insurrection.

Anarchism in the Americas suffered similar reverses. In the United States, a native and mainly nonviolent tradition developed during the 19th century in the writings of Henry David Thoreau, Josiah Warren, Lysander Spooner,

and Benjamin Tucker (editor of *Liberty*, an anarchist journal published in Boston and later in New York City, 1881–1908). Activist Anarchism in the U.S. was mainly sustained by immigrants from Europe, including Johann Most (editor of *Die Freiheit*), Alexander Berkman (who attempted to assasinate steel magnate Henry Clay Frick in 1892), and Emma Goldman, whose *Living My Life* gives a picture of radical activity in the United States at the turn of the century. Anarchism appeared as a dramatic element in American life in 1886, when seven policemen were killed in the Haymarket bombing in Chicago and four Anarchist leaders were executed—unjustly, as later investigations revealed. In 1901 a Polish Anarchist, Leon Czolgosz, assasinated President McKinley. In 1903 the U.S. Congress passed a law to bar foreign Anarchists from the country and to deport alien Anarchists found within it. In the repressive mood that followed World War I the Anarchists, like the IWW, were supressed; Alexander Berkman, Emma Goldman, and many others were imprisoned and deported.

Haymarket bombing

In Latin America strong Anarchist elements were involved in the Mexican Revolution. The Syndicalist teachings of Ricardo Flores Magon influenced the peasant revolutionism of Emiliano Zapata. After the deaths of Zapata in 1919 and Flores Magon in 1922, the revolutionary image in Mexico, as elsewhere, was taken over by Communists. In Argentina and Uruguay considerable Anarcho-Syndicalist movements existed early in the 20th century, but they too were greatly reduced by the end of the 1930s through intermittent repression and the competition of Communism.

CONTEMPORARY ANARCHISM

After World War II, anarchist groups and federations reemerged in almost all countries where they had formerly flourished—the notable exceptions being Spain and the Soviet Union—but these organizations wielded little influence compared to that of the broader movement inspired by earlier ideas. This development is not surprising, since anarchists never stressed the need for organizational continuity, and the cluster of social and moral ideas that are identifiable as anarchism always spread beyond any clearly definable movement.

Anarchist ideas emerged in a wider frame of reference beginning with the American Civil Rights Movement of the 1950s, which aimed to resist injustice through the tactic of civil disobedience. In the 1960s and '70s a new radicalism took root among students and the left in general in the United States, Europe, and Japan, embracing a general criticism of "elitist" power structures and the materialist values of modern industrial societies—both capitalist and communist. For these radicals, who rejected the traditional parties of the left as strongly as they did the existing political structure, the appeal of anarchism was strong. The general anarchist outlook—with its emphasis on spontaneity, theoretical flexibility, simplicity of life, and the importance of love and anger as complementary and necessary components in both social and individual action—attracted those who opposed impersonal political institutions and the calculations of older parties. The anarchist rejection of the state, and the insistence on decentralism and local autonomy, found strong echoes among those who advocated participatory democracy. The anarchist insistence on direct action was reflected in calls for extraparliamentary action and violent confrontation by some student groups in France, the United States, and Japan. And the recurrence of the theme of workers' control of industry in so many manifestos of the 1960s—especially during the student uprisings in Paris in May 1968—showed the enduring relevance of anarcho-syndicalist ideas.

Beginning in the 1970s, anarchism became a significant factor in the radical ecology movement in the United States and Europe. Anarchist ideas in works by the American novelist Edward Abbey, for example, inspired a generation of "eco-anarchism" in the United States, including the radical Earth First! organization, to protest urban sprawl and the destruction of old-growth forests. Much influential work in anarchist theory during this period and afterward, such as that of Murray Bookchin, was noteworthy for its argument that statism and capitalism were incompatible with environmental preservation.

Eco-anarchism

Anarchists also took up issues related to feminism and developed a rich body of work, known as "anarcha-feminism," that applied anarchist principles to the analysis of women's oppression. They argued that the state is inherently patriarchal and that women's experience as nurturers and caregivers reflects the anarchist ideals of mutuality and the rejection of hierarchy and authority.

The most prevalent current in anarchist thinking during the last two decades of the 20th century (at least in the United States) was an eclectic, countercultural mixture of theories strongly influenced by postmodern philosophy and literary theory, in particular by the writings of French philosopher and historian Michel Foucault. From the 1970s a distinct African American anarchist movement, represented in the writings of former Black Panther Lorenzo Kom'boa Ervin, has been influential among anarchists in the United States and other parts of the world.

In 1999 anarchist-led demonstrations against the World Trade Organization in Seattle provoked wide media attention, as did later related protests against the World Bank and the International Monetary Fund. The unprecedented publicity given to the anarchists' explicitly revolutionary viewpoint inspired a proliferation of new anarchist groups, periodicals, and Internet sites. Anarchists were also a significant—and in some cases a predominant—influence in many other political movements, including campaigns against police brutality and capital punishment, the homosexual rights movement, and diverse movements promoting animal rights, vegetarianism, abortion rights, and the legalization of marijuana.

Protests in Seattle

At the beginning of the 21st century, no anarchist movement posed a serious threat to state power, and anarchists were no closer to achieving their dream of a society without government than they were a century before. Nevertheless, for many observers and followers around the world, anarchism remained an active and vibrant ferment of criticism, protest, and direct action.

(G.W./M.M./Fr.R./Ed.)

Fascism

Fascism comprises a political ideology and mass movement that dominated many parts of central, southern, and eastern Europe between 1919 and 1945 and also found adherents in western Europe, the United States, South Africa, Japan, Latin America, and the Middle East. Europe's first fascist leader, Benito Mussolini, took the name of his party from the Latin word *fasces*, which referred to a bundle of elm or birch rods (usually containing an ax) used as a symbol of penal authority in ancient Rome. Although fascist parties and movements differed significantly from each other, they had many characteristics in common, including extreme nationalism, contempt for democracy and liberalism, and the scapegoating of leftists, Freemasons, immigrants, and various ethnic groups, especially Jews.

At the end of World War II, the major European fascist parties were broken up, and in some countries, such as Italy and West Germany, they were officially banned. Beginning in the late 1940s, however, many "neofascist" parties and movements were founded in Europe as well as in Latin America and South Africa.

NATIONAL FASCISMS

Fascist parties and movements arose in several European countries between 1922 and 1945. Among the most significant were the National Fascist Party (Partito Nazionale Fascista) in Italy, led by Mussolini; the National Socialist German Workers' Party (Nationalsozialistische Deutsche Arbeiterpartei), or Nazi Party, (Nationalsozialistische Deutsche Arbeiterpartei), led by Adolf Hitler and representing his National Socialism movement; the Fatherland Front (Vaterländische Front) in Austria, led by Engelbert Dollfuss and supported by the Heimwehr (Home Defense Force); the Falange ("Phalanx") in Spain, founded in 1933 by José Antonio Primo de Rivera, many of whose members were absorbed into the military dictatorship of Francisco Franco; the Cross of Fire (Croix de Feu), later renamed the French Social Party (Parti Social Française), led by Colonel François de La Rocque; and the National

Union (Nasjonal Samling) in Norway, whose leader, Vidkun Quisling, was made minister president under the German occupation.

Some important fascist movements outside Europe during this period were the military dictatorship of Admiral Tojo Hideki in Japan and, in South Africa, the Gentile National Socialist Movement and its splinter group, the South African Fascists. In the United States the Ku Klux Klan, a white supremacist organization founded at the end of the Civil War and revived in 1915, displayed some fascist characteristics.

COMMON CHARACTERISTICS OF FASCISM

There has been considerable disagreement among historians and political scientists about the nature of fascism. Some scholars, for example, regard it as a socially radical movement with ideological ties to the Jacobins of the French Revolution, whereas others see it as an extreme form of conservatism inspired by the 19th-century reaction against the ideals of the Enlightenment. Some find fascism deeply irrational, whereas others are impressed with the rationality with which it served the material interests of its supporters. Some consider fascism to be motivated primarily by its aspirations—by a desire for cultural "regeneration" and the creation of a "new man"—others place greater weight on fascism's "anxieties"—on its fear of communist revolution and even of left-centrist electoral victories.

Difficulties of defining fascism

One reason for these disagreements is that the two historical regimes that are today regarded as paradigmatically fascist—Mussolini's Italy and Nazi Germany—were different in important respects. In Italy, for example, anti-Semitism was officially rejected until 1938, when Mussolini abruptly enacted a series of anti-Semitic measures in order to solidify his new military alliance with Hitler. Another reason is the fascists' well-known opportunism—*i.e.,* their willingness to make changes in official party positions in order to win elections or consolidate power. Finally, scholars of fascism themselves bring to their studies different political and cultural attitudes, which often have a bearing on the importance they assign to one or another aspect of fascist ideology or practice. Secular liberals, for example, have stressed fascism's religious roots; Roman Catholic and Protestant scholars have emphasized its secular origins; social conservatives have pointed to its "socialist" and "populist" aspects; and social radicals have noted its defense of "capitalism" and "elitism."

For these and other reasons, there is no universally accepted definition of fascism. Nevertheless, it is possible to identify a number of general characteristics that fascist movements of the interwar period tended to have in common. Among them were: extreme nationalism; totalitarianism and the rejection of Marxism, parliamentary democracy, and liberalism; belief in the *Führerprinzip,* or "leadership principle"; the pervasive use of violence; support of economic elites; imperialism; the attempt to create a "new man"; an obsession with "decadence" and "spirituality"; and scapegoating of groups such as Jews, leftists, and immigrants.

Extreme nationalism. Whereas cosmopolitan conservatives often supported international cooperation and admired elite culture in other countries, fascists espoused extreme nationalism and cultural parochialism. Fascist ideologues taught that national identity was the foundation of individual identity and should not be corrupted by foreign influences, especially if they were left-wing. Fascists in general wanted to replace internationalist class solidarity with nationalist class collaboration, and they regularly accused their political opponents of being less "patriotic" than they, sometimes labelling them as "traitors."

Fascist nationalism in France was particularly hostile to immigrants, especially if they were leftists. Jean Renaud of French Solidarity (Solidarité Française) demanded that all foreigners seeking residence in France be rigorously screened and that the unfit be denied entry "without pity"—especially, social revolutionaries, who, he said, made France "not a refuge for the oppressed but a despository for trash."

Totalitarianism. As Führer ("Leader") of the Third Reich, Hitler attempted not only to control all political power but also to dominate many institutions and organizations that previously were independent of the state, such as courts, churches, universities, social clubs, veterans' groups, sports associations, and youth groups. Even the German family came under assault, as members of the Hitler Youth were instructed to inform on anti-Nazi parents. In Italy, Mussolini adopted the title of Duce ("Leader"), and his regime created billboards displaying slogans such as "Il Duce ha sempre ragione" (The Duce is always right") and "Credere, obbedire, combattere" ("Believe, obey, fight").

Rejection of Marxism. Fascists made no secret of their hatred of Marxists of all stripes, from totalitarian communists to democratic socialists, and they promised to deal more "firmly" with Marxists than had earlier, more democratic rightist parties. Mussolini first made his reputation as a fascist by unleashing armed squads of Black Shirts on striking workers and peasants in 1919–20. Many early Nazis had served in the Freikorps, the paramilitary groups of ex-soldiers formed to suppress leftist activism in Germany at the end of World War I. The Nazi SA (Sturmabteilung ["Assault Division"], or Brownshirts) clashed regularly with German leftists in the streets before 1933, and when Hitler came to power he sent hundreds of Marxists to concentration camps and intimidated "red" neighbourhoods with police raids and beatings.

In 1919–20 the Heimwehr in Austria performed the same function that the Freikorps did in Germany, its volunteer militia units doing battle with Marxists and perceived foreign enemies. Many of these units were organized by members of the landed gentry and the middle class to counter strikes by workers in the industrial districts of Linz and Steyer. In 1927 violent clashes between the Heimwehr and the Schutzbund, a socialist defense organization, resulted in many deaths and injuries among the leftists. In 1934 the Heimwehr joined Dollfuss's Fatherland Front and was instrumental in pushing Dollfuss toward fascism.

In 1919 a number of fascist groups emerged in Japan to resist new demands for democracy and to counter the influence of the Russian Revolution of 1917. They frequently acted as strikebreakers; launched violent assaults on left-wing labour unions, peasant unions, and the socialist Levelling Society; and disrupted May Day celebrations.

Fascism and Soviet communism

Despite the fascists' violent opposition to Marxism, some observers have noted significant similarities between fascism and Soviet communism. Both were mass movements, both emerged in the years following World War I in circumstances of political turmoil and economic collapse, both created totalitarian systems after they came to power (and often concealed their totalitarian ambitions beforehand), and both employed terror and violence without scruple when it was expedient to do so. Other scholars have cautioned against reading too much into these similarities, however, noting that fascist regimes (in particular Nazi Germany) used terror for different purposes and against different groups than did the Soviets, and that fascists, unlike communists, generally supported capitalism and defended the interests of economic elites.

Rejection of parliamentary democracy. Fascist movements criticized parliamentary democracy for allowing Marxism to exist at all. According to Hitler, democracy undermined the natural selection of ruling elites and was "nothing other than the systematic cultivation of human failure." Similarly, the Heimwehr declared in 1930 that it would overcome class struggle by replacing political democracy with "a strong national leadership which will consist not of the representatives of parties but of leading members of the large corporations and of the ablest, most trustworthy men in our own mass movement." In 1935 La Rocque condemned elections as exercises in "collective decadence," and early in 1936 he told his followers that "even the idea of soliciting a vote nauseates me." The Tojo dictatorship in Japan dissolved all political parties, even those on the right, and reduced other political freedoms.

Before coming to power, fascists sometimes engaged in electoral politics to give the appearance of submitting to democratic procedures. When Hitler was appointed chancellor in 1933, he abandoned his military uniform for a civilian suit and bowed profusely to President Paul von Hindenburg in public ceremonies. In 1936, faced with the

prospect that the Cross of Fire would be banned by the government as a paramilitary organization, La Rocque founded a new and ostensibly more democratic party, the French Social Party, which he publicly claimed was "firmly attached to republican liberties." However, he privately made it clear to his followers that his conversion was tactical rather than principled.

Rejection of political and cultural liberalism. Although circumstances sometimes made accommodation to political liberalism necessary, fascists condemned this doctrine for placing the rights of the individual above the needs of the *Volk* ("people"), encouraging "divisiveness" (*i.e.*, political pluralism), tolerating "decadent" values, and limiting the power of the state, and they accused liberal "fellow travelers" of wittingly or unwittingly abetting communism.

Fascist propagandists also attacked cultural liberalism, claiming that it encouraged moral relativism, godless materialism, and selfish individualism and thereby undermined traditional morality. Anti-Semitic fascists associated liberalism with Jews—indeed, one precursor of Nazism, the political theorist Theodor Fritsch, claimed that to succumb to a liberal idea was to succumb to the Jew within oneself.

"The Jew within oneself"

Violence. Fascists preferred to crush their political opponents with physical force. Before he came to power, Mussolini sent his Blackshirts to assault socialist organizers throughout Italy; Hitler's storm troopers served a similar function. In 1931–36 Japanese fascists assassinated a number of important political figures and business leaders, and in the United States in the 1920s and '30s the Ku Klux Klan and other groups sought to intimidate African Americans with cross-burnings, beatings, and lynchings.

The Führerprinzip. Fascists defended the *Führerprinzip*, the belief that the party and the state should be led by one person with absolute power. This principle was also applied at lower levels of the political and social hierarchy, reflecting a so-called "corporal syndrome" in which persons willingly submit to the authority of those above them in exchange for the gratification they derive from dominating those below. Japanese fascists, for example, believed that owners of stores and workshops should exercise "paternal" authority over their workers, who should not be permitted to form unions. Such "small bosses" assumed the leadership of town and village councils throughout Japan.

Support of economic elites. The economic programs of the great majority of fascist movements were extremely conservative, favouring the wealthy far more than the middle class and the working class. Nazi "anticapitalism," such as it was, was aimed primarily at Jews; non-Jewish capitalists were allowed to keep their companies and their wealth, following a distinction made in the Nazi Party's original program and never changed. Mussolini, a leading member of the Italian Socialist Party (Partito Socialista Italiano; PSI) before World War I, became a fierce antisocialist after the war. Until he instituted a war economy in the mid-1930s, he allowed industrialists to run their companies with a minimum of government interference, and he cut taxes on business, permitted cartel growth, decreed wage reduction, and rescinded the eight-hour workday law. Between 1928 and 1932, real wages in Italy dropped by almost half. Mussolini admitted that the standard of living had fallen but stated that "fortunately the Italian people were not accustomed to eating much and therefore feel the privation less acutely than others."

Imperialism. Many fascist movements had imperialistic aims. Hitler hoped that his *Drang nach Osten* ("drive toward the east"), by conquering eastern Europe and Russia, would not only prove the racial superiority of Aryans over Slavs but also provide enough plunder and *Lebensraum* ("living space") to overcome continuing economic difficulties at home. Mussolini's imperial ambitions were directed at North Africa, and his armies invaded Ethiopia in 1935. Japanese fascists preached military conquest in keeping with their plan for a "Greater East Asia Co-Prosperity Sphere." French fascists were strong defenders of the French empire in Indochina and North Africa, and during the interwar period they attracted considerable support among the ruling European minority (*colons*) in Algeria.

Lebensraum

The "new man." Fascists aimed to transform the ordinary man into the "new man," a "virile" being who would put decadent bourgeoisie, cerebral Marxists, and "feminine" liberals to shame. The new man, by contrast, would be physically strong and morally "hard." As Hitler described him, the new man was "slim and slender, quick like a greyhound, tough like leather, and hard like Krupp steel." The French fascist writer Pierre Drieu La Rochelle claimed that the Hitlerian man "is a type who rejects culture, who stands firm in the middle of sexual and alcoholic depravity and who dreams of bringing to the world a physical discipline with radical effects." The new man was also a Darwinian "realist" who was contemptuous of "delicate" souls who refused to employ harsh military or political measures when they were required.

During World War II, Heinrich Himmler, chief of the SS (Schutzstaffel ["Protective Echelon"]), the elite military police of the Nazi Party, addressed an SS unit that had executed many Jews, reminding his "new men" of the need to be emotionally hard: "Most of you know what it means when 100 corpses are piled up, when 500 or 1,000 are piled there. To have gone through this and . . . to have remained decent, that is what has made us hard. I have to expect of you superhuman acts of inhumanity. . . . We have no right to be weak. . . . [Our men] must never be soft. They must grit their teeth and do their duty."

Decadence and spirituality. Some of the ugliest aspects of fascism were fueled by what fascists saw as a morally justified struggle against "decadence." For fascists, decadence meant a number of things: materialism, self-indulgence, hedonism, cowardice, and physical and moral softness. It was also associated with rationalism, skepticism, atheism, humanitarianism, and democracy, as well as with rule by the Darwinian unfit, including the weak and the "female."

The opposite of decadence was "spirituality," which transcended materialism and generated self-discipline and virility. The spiritual attitude involved a certain emotional asceticism that enabled one to avoid feelings of pity for one's victims. It also involved Darwinian notions of survival of the fittest, a belief in the right of natural elites to upward social and political mobility, and accommodation with the upper classes. It prized hierarchy, respect for superiors, and military obedience. It was forceful toward the weak, and it was "male." The spiritual attitude was also hateful. In 1934 Ernst Röhm, leader of the SA, worried that Germans had "forgotten how to hate." "Virile hate," he wrote, "has been replaced by feminine lamentation."

Scapegoating. Fascists often blamed their countries' problems on scapegoats. Jews, Freemasons, Marxists, and immigrants were prominent among the groups that were demonized. According to fascist propaganda, the long depression of the 1930s resulted less from insufficient government regulation of the economy or inadequate lower-class purchasing power than from "judeo-masonic-bolshevik" conspiracies, left-wing agitation, and the presence of immigrants.

VARIETIES OF FASCISM

Just as Marxists, liberals, and conservatives differed within and between various countries, so too did fascists. In some countries there were rivalries between native fascist movements over personal, tactical, and other differences. Fascist movements also displayed significant differences with respect to their acceptance of racism (particularly anti-Semitism), their identification with Christianity, and their support for Nazi Germany.

Acceptance of racism. Although not all fascists believed in racism, it played a central role in the actions of those who did. Nazism was viciously racist, especially in its attitude toward Jews. The Nazis blamed the Jews for almost everything wrong with Germany, from the Great Depression and the rise of Marxism to the evils of international capitalism and decadence in art. The Holocaust, culminating in the "final solution to the Jewish question," was the immensely cruel outcome of this hatred. From 1933 to 1945 some six million Jewish men, women, and children were exterminated by gassings, shootings, hangings, and clubbings, and about three million Slavs, as well as approximately 400,000 Gypsies (Roma), were murdered.

"The final solution"

Croatian fascists preached the racial inferiority of Serbs, and in the late 1930s they became increasingly anti-Semit-

ic. When Germany invaded Yugoslavia in 1941, Ante Pavelič, leader of the Croatian Ustaša ("Insurgence"), became head of a German puppet state, the Independent State of Croatia, and established a one-party regime. Pavelič forced some of the more than one million Orthodox Serbs in Croatia to convert and expelled or killed others in campaigns of genocide. About 250,000 Serbs in Croatia were eventually liquidated; the regime also murdered some 40,000 Jews.

Elsewhere in Europe and in South Africa, Latin America, and the United States, fascist movements were racist, and sometimes specifically anti-Semitic, to varying degrees. In contrast to fascists in most other European countries, the Italian Fascists were opposed to anti-Semitism during the first 14 years of their rule. In 1938, however, the Italian government passed anti-Semitic legislation, and later it abetted the Holocaust.

During the early interwar period, France's largest fascist parties rejected anti-Semitism, and right-wing Jews were accepted into these movements until at least 1936, when the left-wing Popular Front, under the premiership of the Jewish socialist Léon Blum, came to power. Other fascist groups, such as French Action and French Solidarity, were more openly anti-Semitic, though they claimed to object to Jews on "cultural" rather than racial grounds.

Identification with Christianity. Most fascist movements portrayed themselves as defenders of Christianity against atheists and amoral humanists. This was especially true of the Falanga movement in Poland, which defended ultramontane Catholicism, and the Falange in Spain. In 1933–34, a number of Protestant periodicals in Prussia promoted Hitler as a protector of the Christian family against cultural modernism. The Nazis themselves were rarely atheists, and most German anti-Semites supported Christianity purged of its "Jewish" elements. The pro-Nazi "German Christians," part of the Lutheran church in Germany, called themselves "SS men for Christ" and insisted that Christ had been a blond-haired, blue-eyed Aryan.

"SS men for Christ"

Although fascists in Germany and Italy posed as protectors of the church, their ideologies contained many elements that conflicted with traditional Christian beliefs. For example, the Nazis rejected the Christian ideals of humility and mercy on the grounds that they repressed the violent instincts necessary to prevent inferior races from dominating Aryans. Martin Bormann, the second-most powerful official in the Nazi Party after 1941, argued that Nazi and Christian beliefs were "incompatible," primarily because the essential elements of Christianity were "taken over from Judaism." Bormann's views were shared by Hitler, who ultimately wished to replace Christianity with a racist form of warrior paganism.

In 1929 Mussolini signed a concordat with the papacy, the Lateran Treaty, which made Roman Catholicism the state religion of Italy. In 1931, however, Pope Pius XI issued an encyclical, *Non abbiamo bisogno,* that denounced fascism's "pagan worship of the state."

Support for Germany. Fascist nationalism led many non-German fascists to oppose Nazi Germany. Many Polish fascists, for example, were killed while resisting the German invasion of Poland in 1939, and others were later condemned to Nazi concentration camps. Before 1940, all French fascists opposed a German invasion of France. Following the country's defeat, some French fascists, such as Philippe Barrès, crossed the channel to serve under Charles de Gaulle, leader of the Free French movement. Italy and Japan were allies of Germany during the war, though Mussolini's autonomy in the alliance was lost when German divisions occupied the country in 1942.

ORIGINS AND SUPPORT

Mussolini and Hitler did not invent fascist ideology. Many fascist ideas derived from the reactionary backlash to the progressive revolutions of the late-18th and 19th centuries and to the secular liberalism and social radicalism that accompanied these upheavals. The political theorist Joseph de Maistre condemned the Enlightenment for subverting the dominance of religion and traditional elites, and he paid homage to the public executioner as the protector of a divinely sanctioned social hierarchy. The historian Hippolyte Taine lamented the rise to power of the masses, whom he suggested were at a lower stage of biological evolution than aristocrats. Maurice Barrès, a French writer and politician, fused ethnic rootedness with authoritarian nationalism and contended that too much civilization led to decadence and that hatred and violence were energizing remedies.

Joseph de Maistre

German populist politicians and writers such as Adolf Stoecker, Otto Böckel, and Theodor Fritsch extolled the idea of racially pure peasants close to the soil who would one day follow a charismatic leader able to intuit the *Volk* soul without benefit of elections. Anti-Semitism was a staple in the work of many best-selling authors in Germany and France, and several anti-Semitic ideas espoused by Carl Lueger's Christian Social Party and Georg von Schönerer's Pan-German movement in Austria were later adopted by Hitler.

Despite their long history in European thought, fascist ideas prospered politically only when economic threats increased their appeal to members of certain social groups. In 1928, before the onset of the Great Depression in Germany, Hitler received less than 3 percent of the vote; after 1930, far more voters—many of them middle- and lower-middle-class workers fearful of "proletarianization"—gave him their support. The economic anxiety underlying the success of Nazism was reflected to some extent in party membership, which was drawn disproportionately from economic elites and other high-status groups. Similarly, in Italy, fascism originally received most of its support from large and small landowners, businessmen, and white-collar workers.

Because they shared some of the fascists' goals, some nonfascist conservatives were led to collaborate with fascists in times of crisis. During the Great Depression, for example, thousands of middle-class conservatives fearful of the growing power of the left abandoned traditional right-wing parties and adopted fascism. The ideological distance traveled from traditional conservatism to Nazism was sometimes small, since many of the ideas that Hitler exploited in the 1930s had long been common currency within the German right. In Italy thousands of landowners and businessmen were grateful to Mussolini's Blackshirts for curbing the socialists in 1922, and many in the army and the Catholic church saw fascism as a bulwark against communism.

Fascists also received support from Christian conservatives. Between 1930 and 1932 Hitler was popular among many Protestant voters in rural Prussia, and after 1933 the Catholic church in Germany largely accommodated itself to his regime. In France the leading Catholic newspaper, *La Croix,* expressed early support for Hitler's crusade against bolshevism, and the largest Catholic parliamentary party, the Republican Federation (Fédération Republicaine), included many fascists in its ranks.

NEOFASCISM

Although fascism was largely discredited in Europe at the end of World War II, fascist-inspired movements were founded in several European countries beginning in the late 1940s. Similar groups were created outside Europe as well, primarily in Latin America, the Middle East, and South Africa. Like their predecessors, the "neofascists" advocated militant nationalism and authoritarianism, opposed the liberal individualism of the Enlightenment, attacked Marxist and other left-wing ideologies, indulged in racist and xenophobic scapegoating, portrayed themselves as protectors of traditional national culture and religion, glorified violence and military heroism, and promoted populist right-wing economic programs.

Despite these similarities, neofascism was not simply a revival of fascism. Neofascist parties differed from earlier fascist movements in several significant respects, many of them having to do with the profound political, economic, and social changes that took place in Europe in the first decades after the end of the war. For example, neofascists tended to place more blame for their countries' problems on non-European immigrants than on leftists and Jews. Also, after decades of postwar decolonization, neofascists in western Europe had little interest in taking *Lebensraum*

Non-European immigrants

through military conquest of other states. Instead, they fought battles for "urban space," which in Germany involved conflicts over government-subsidized housing for immigrants. With increasing urbanization also came a shift in the electoral bases of fascist-oriented movements and a consequent decline in the importance of rural romanticism ("blood and soil") in neofascist political rhetoric. Finally, the gradual acceptance of democratic norms by the vast majority of western Europeans reduced the appeal of authoritarian ideologies and required that neofascist parties make a concerted effort to portray themselves as democratic and "mainstream."

As with fascist movements of the interwar period, neofascists differed from one another in various respects. The rhetoric of neofascists in Russia and the Balkans, for example, tended to be more openly brutal and militaristic than that of the majority of their Western counterparts. Most neofascist movements in Europe pandered to anti-Semitism, though neofascists in Italy and Spain generally did not. Spanish neofascists also differed from most other neofascists in Europe in that they did not make a major issue of immigration. In the 1990s in Russia and eastern Europe, neofascist movements were generally more leftist than their counterparts in western Europe, emphasizing the interests of workers and peasants over those of the urban middle class and calling for "mixed" socialist and capitalist economies.

National movements. *Italy.* The Italian Social Movement

One of the largest neofascist parties in western Europe in the 1990s was the Italian Social Movement (Movimento Sociale Italiano [MSI]; renamed the National Alliance [Alleanza Nazionale; AN] in 1994). Founded in 1946, it was led at various times by Giorgio Almirante, Augusto De Marsanich, Arturo Michelini, and Gianfranco Fini.

Although Italy's postwar constitution forbade the reorganization of a fascist party, the propaganda of the MSI echoed a number of themes dear to interwar fascism. Foremost was its call for the "vital forces" of the nation to resist the communist menace. In the 1950s MSI members entered schools to assault leftists and provoked violent confrontations with socialist and communist activists during election campaigns and strikes.

MSI electoral fortunes varied greatly according to circumstances, ranging from about 2 percent of the vote in 1948 to 13.5 percent in 1994. In local elections in 1993 Fini and MSI member Alessandra Mussolini, the granddaughter of the Duce, were nearly elected mayors of Rome and of Naples, respectively.

Immediately following these elections, Fini subsumed the MSI into a new party, the National Alliance. Officially rejecting "any form of dictatorship or totalitarianism," he replaced the old slogan of a "third way" between capitalism and communism with praise for the free market and individual initiative. In March 1995 the AN won about 14 percent of the vote and five ministerial posts in a coalition government led by Silvio Berlusconi. Later that year the AN led an attempt to repeal the clause in the Italian constitution forbidding the reorganization of a fascist party, but the effort failed. Although Fini described the AN as "post-fascist," following the 1994 elections he declared that Mussolini was the greatest Italian statesman of the 20th century and that fascism before 1938 (*i.e.*, before Mussolini formed a military alliance with Hitler) was "mostly good." In 2001 the AN joined a new coalition government under Berlusconi.

Germany. In 1949 Fritz Dorlis and Otto Ernst Remer, a former army general who had helped to crush an attempted military coup against Hitler in July 1944, founded the Socialist Reich Party (Sozialistische Reichspartei; SRP), one of the earliest neofascist parties in Germany. Openly sympathetic to Nazism, the SRP made considerable gains in former Nazi strongholds. It was banned as a neo-Nazi organization in 1952.

Among legal neofascist parties in Germany, the most important were the National Democratic Party of Germany (Nationaldemokratische Partei Deutschlands; NPD), founded in 1964 by Waldemar Schütz, a former member of Waffen-SS (the elite military wing of the Nazi Party); the German People's Union (Deutsche Volksunion; DVU), founded in 1971; and the Republicans (Die Republikaner; REP), founded in 1983 by another former Waffen-SS member, Franz Schönhuber. Die Republikaner

Neofascist parties in Germany focused much of their energies on campaigns against immigrants, and they were most successful in areas where immigrant communities were large. They also won significant support among disaffected youth in parts of the former East Germany, which was plagued by high levels of unemployment, poor housing, and severe environmental problems in the years immediately following unification.

In 1992–93 gangs of neo-Nazi youth in eastern Germany staged attacks on Turkish and other immigrants and desecrated Jewish cemeteries. Public revulsion at the attacks contributed to a temporary dip in the far-right vote in 1993. At the end of the 1990s the REP was torn by personal, generational, and tactical divisions, with some members favouring a blatantly pro-Nazi platform and others urging more moderate and mainstream positions.

Austria. In 1999–2000 a series of electoral successes by the far-right Freedom Party of Austria (Freiheitlichen Partei Österreichs; FPÖ), founded in 1956 and led from 1986 by Jörg Haider, created a storm of controversy and produced widespread protests in Austria and abroad, largely because of perceptions that the leadership of the party, including Haider himself, was sympathetic to Nazism. Haider, whose father had been a leading member of the Austrian Nazi Party before and during World War II, became notorious for his praise of Hitler's employment policies and his remark, made to a group of Austrian veterans of World War II, that the Waffen-SS deserved "honour and respect." Arguing for stricter controls on immigration, he warned against the "over-foreignization" of Austrian society, pointedly borrowing a term—*Überfremdung*—used by Joseph Goebbels, Hitler's minister of propaganda.

In general elections in October 1999, the FPÖ narrowly outpolled the conservative Austrian People's Party (Österreichische Volkspartei; ÖVP) with 27 percent of the vote, thereby becoming the second largest party in Austria (the Social-Democratic Party of Austria [Sozialdemokratische Partei Österreichs; SPÖ] finished first, with more than 33 percent). The ÖVP, with considerable reluctance, formed a government with the FPÖ in February 2000, prompting the European Union (EU) to carry out its threat to suspend bilateral political contacts with Austria. The new government was greeted by widespread demonstrations, diplomatic protests, and calls for boycotts on Austrian tourist destinations. Facing intense international pressure, Haider resigned his leadership of the FPÖ at the end of February, only three weeks after his party entered the government.

France. The National Front

In the 1980s and '90s neofascism in France was dominated by the National Front (Front National; FN), founded in 1972 by François Duprat and François Brigneau and led from that year by Jean-Marie Le Pen. After 10 years on the margins of French politics, the FN began a period of spectacular growth in 1981. Campaigning on the slogan "France for the French" (as had French fascists in the 1930s) and linking high unemployment and increased crime to the presence of immigrants, the FN increased its support from 1 percent of the vote in 1981 to 14 percent in 1988. The party's anti-immigrant themes also included the claim that non-French immigrants, especially Muslims, threatened French national identity and culture and that liberal permissiveness and multiculturalism had eroded traditional values. In 1984 the FN gained 11 percent of the vote in elections for the European Parliament, thereby becoming the largest extreme-right group within that body. Le Pen won 15 percent of the vote in presidential elections in 1995, and the FN also took 15 percent in legislative elections in May–June 1997.

The FN's rapid growth occurred despite Le Pen's previous association with extreme right-wing causes, his cavalier remarks about the Holocaust (in 1987 he told a television interviewer that the Holocaust was only "a detail of history"), the presence of former fascists in his organization, and other neofascist aspects of his movement. Although Le Pen described himself as a "Churchillian democrat," his commitment to political democracy was similar to that of

La Rocque in the 1930s and '40s—more tactical than principled. "We must be respectful of legality while it exists," he said in 1982. Le Pen admired Francisco Franco, and he praised Augusto Pinochet's overthrow of Chilean socialist president Salvador Allende in 1973, declaring that the French army should follow Pinochet's example if a similar leftist government were to arise in France.

The FN attempted to portray Le Pen as a plain-speaking man of the people, and it emphasized his physical strength and virility. Although Le Pen's bodyguards sometimes wore helmets and battle gear similar to those of France's national riot police, and although party supporters were sometimes involved in street violence against immigrants and ethnic minorities, the FN had no official party uniforms or paramilitary organizations.

The FN imposed censorship when it had the power to do so. Mayors of cities governed by the FN removed left-wing journals from municipal libraries, forbade librarians from ordering "internationalist" books, and required the purchase of materials supporting the FN's views. The mayor of Toulon, Jean-Marie Le Chevallier, canceled the award of a literary prize to a Jewish writer and tried to shut down a well-known performance festival in the city because of its leftist political orientation.

By the 1990s the FN had acquired a broad-based and diverse following, including small businessmen and self-employed artisans, unemployed white-collar and blue-collar workers, socially conservative Catholics, and young people. In 1998 Le Pen's associate Bruno Mégret split from the FN to form a new party, the National Movement (Mouvement National; MN), taking with him most of the FN's departmental secretaries and city councillors. For Le Pen, Mégret's action was a "crime against France." In elections for the European Parliament in 1999, the two parties received a total of only 9 percent of the vote, a major setback for Le Pen and French neofascism.

Russia. After World War II few Russians needed to be reminded of the evils of German fascism. Nevertheless, several fascist groups emerged in Russia after the breakup of the Soviet Union in 1991. Resentment over the loss of the Soviet empire, concern for the fate of ethnic Russians in the successor states, bad economic conditions, the breakdown of law and order, the desire for a strong leader, and the fact that democratic institutions were not deeply rooted in Russia—all combined to make fascist ideas appealing to some segments of the Russian population.

The Black Hundreds

Some Russian fascists attempted to revive the reactionary ideology of the Black Hundreds, a loose association of extreme right-wing organizations formed in Russia during the early years of the 20th century. Black Hundred ideology was highly nationalistic, anticosmopolitan, anti-Semitic, anti-Masonic, anti-Western, antidemocratic, antiegalitarian, antiliberal, and anti-"decadence." The Black Hundreds were strong supporters of the Russian Orthodox church, the army, and authoritarian government, and they indulged in conspiracy theories that blamed most of Russia's troubles on Jews and Freemasons.

In the 1990s the leading group espousing Black Hundred ideology was Pamyat ("Memory"), whose main spokesman after 1984 was Dmitry Vasiliev. Pamyat writers denounced communists as "godless," "cosmopolitan," and "antipatriotic," and they criticized the neglect of national traditions, anti-Russian sentiment in the Baltic countries, the moral decline of youth, increased crime, the weakening of the family, and alcoholism. Although Pamyat had a near monopoly on the extreme right in 1987–88, by 1991 it had been overtaken by rival movements.

One of these movements was the Liberal-Democratic Party of Russia (Liberalno-Demokraticheskaya Partiya Rossi; LDPR), led by Vladimir Zhirinovsky. Founded in 1990, the party grew rapidly, and in presidential elections in 1991 Zhirinovsky won almost 8 percent of the vote, placing him third after Boris Yeltsin and Nicolay Ryzhkov. In parliamentary elections in 1993 the LDPR gained nearly 23 percent of the vote, more than did the Russian Communist Party (12.4 percent). However, by 1996 Zhirinovsky's support had declined precipitously, and he won only 6 percent of the vote in presidential elections in that year and less than 3 percent in 2000.

Zhirinovsky and "Russian fascism"

Most neofascists denied that they were fascists, and Zhirinovsky was no exception. On various occasions he declared his adherence to democratic values, the rights of man, a multiparty system, and the rule of law. However, in 1991 he declared: "I say quite plainly, when I come to power there will be a dictatorship. Russia needs a dictator now." He added: "I'll be ruthless. I will close down the newspapers one after another. I may have to shoot 100,000 people, but the other 300 million will live peacefully. You want to call it Russian fascism, fine." Zhirinovsky also indulged in racism and anti-Semitism, even though his own father was apparently Jewish and he himself had been active in a Russian Jewish group in 1989.

Zhirinovsky wanted to ensure Russia's greatness by retaining control of the constituent republics of the former Soviet Union, and he condemned independence movements in the Baltic states and Chechnya and threatened harsh measures against them. As he told a Lithuanian newspaper in 1991, "I'll destroy you. I'll bury nuclear waste . . . along the border [with the Baltic states]. . . . You Lithuanians will die from diseases and radiation. . . . Soon there will be no Lithuanians, Estonians, and Latvians in the Baltic. I'll act the way Hitler did in 1942." Zhirinovsky made similar threats to Western countries, which he believed were working against Russia's interests.

Like many fascists of the interwar period, Zhirinovsky had little regard for women, and he was openly contemptuous of women with education or political power. His opinions were consistent with the negative portrayal of women—especially younger women—in Black Hundreds literature.

Zhirinovsky's economic program favoured a mixed economy. He proposed both that taxes on industry be reduced and that 70 percent of the economy be controlled by the state, including transportation and communication. However, he blamed most of Russia's economic problems on scapegoats, claiming that Russia was so poor because the country had been robbed of its natural resources by Jews, Freemasons, and Americans.

The Russian National Unity (Russkoe Natsionalnoe Edinstvo; RNE), a paramilitary organization founded in 1990 by Aleksandr Barkashov, claimed to have an extensive network of local branches, but its electoral support was significantly less than that of the LDPR. Barkashov, a former commando in the Russian army, touted his blackshirts as a reserve force for the Russian army and the Ministry of Internal Affairs. He blamed many of Russia's economic problems on Jews, insisted that only a "few hundred" Jews had perished in German concentration camps, and said that the Holocaust was a "diversion" created to conceal a Jewish-inspired genocide of 100 million Russians. The RNE's symbol was a left-pointed swastika together with a four-pointed star. The RNE emphasized the "primary importance" of Russian blood, accused "internationalists-communists" of undermining the "genetic purity" of the nation with a program of racial mixture, and called for a rebirth of "Russian-Aryan traditions." Although Barkashov denied that he was a fascist, he admired Hitler enormously, once stating that "I consider [Hitler] a great hero of the German nation and of all white races. He succeeded in inspiring the entire nation to fight against degradation and the washing away of national values."

Ethnic cleansing in the Balkans

Serbia. Following the collapse of communism in the former Yugoslavia and the secession of Croatia and Bosnia and Herzegovina from the Yugoslav federation in 1991–92, units of the Yugoslav army and Serbian paramilitary forces engaged in campaigns of "ethnic cleansing" aimed at driving out non-Serb majorities in northeastern Croatia and parts of northern and eastern Bosnia and establishing nominally independent Serb republics in the vacated territories. The attacks, which were compared in their ferocity and cruelty to the Nazi invasions of eastern Europe and Russia, involved mass executions (mostly of men and boys), forced marches, torture, starvation, and systematic rape. These tactics were aimed at creating irreversible ethnic hatreds that would permanently prevent the development of multiethnic states in the areas under attack. In 1998–99 similar tactics were employed in Kosovo, a province of Serbia in which 90 percent of the population was ethnically Albanian and predominantly Muslim.

Organized and directed by the regime of Serbian president Slobodan Milosevic, leader of the Socialist Party of Serbia (Socijalisticka Partija Srbije; SPS) and later president of Yugoslavia (1997–2000), the campaigns in Croatia and Bosnia were undertaken in part to bolster Milosevic's image as a staunch nationalist and to consolidate his power at the expense of the Vojislav Seselj's Serbian Radical Party (Srpska Radikalna Stranka; SRS), then the largest neofascist party in Serbia. Although the SPS had won 65 percent of the vote in elections to the Serbian assembly in 1990, deteriorating economic conditions and perceived threats to Serbian enclaves in Croatia and Bosnia resulted in a significant loss of support for the SPS and a corresponding growth in the SRS and other extreme nationalist and neofascist groups. To counter the threat from the right, Milosevic gradually adopted many of the neofascists' policies, including support for the creation of a "Greater Serbia" that would incorporate Montenegro, Macedonia, and large areas of Croatia and Bosnia.

In May 1993, after a year of severe economic hardship caused by UN-imposed sanctions, Milosevic accepted an international agreement for the division of Bosnia into 10 ethnic cantons. The plan was rejected by the self-styled parliament of the Bosnian Serbs and condemned by Seselj. Milosevic subsequently attempted to weaken nationalist support for the SRS by allying himself with the notorious paramilitary leader Zeljko Raznjatovic (popularly known by his nom de guerre, Arkan). In elections in December 1993, the SPS increased its representation in the Serbian assembly, taking 49 percent of the vote compared with the SRS's 14 percent.

In early 1998 Serbian military and police forces began attacks in Kosovo on alleged strongholds of the Kosovo Liberation Army (KLA), an ethnically Albanian guerrilla movement fighting to end Serbian control of the province. The Serbs' harsh repression of the Albanian civilian population drew international condemnation and resulted in renewed UN sanctions on Yugoslavia. On March 24, 1999, after a Serbian delegation at peace talks in Rambouillet, France, rejected an accord that had been signed by representatives of Kosovar Albanians and the KLA, the North Atlantic Treaty Organization (NATO) began an intensive bombing campaign directed at Yugoslav military targets and later also at civilian infrastructure and government buildings in Serbia. In response, Serbian security forces in Kosovo conducted a massive campaign of ethnic cleansing, including large-scale massacres of civilians, and eventually forced more than 850,000 Kosovars to flee to border areas in Albania, Macedonia, and Montenegro. The bombing came to an end in early June after Milosevic agreed to the withdrawal of Serbian forces from Kosovo, the deployment of NATO peacekeeping troops, and the repatriation of Albanian refugees. In the meantime, Milosevic and four top officials of his government were indicted for crimes against humanity by the UN International Criminal Tribunal at The Hague. In federal elections held in September 2000, Milosevic was defeated by opposition leader Vojislav Köstunica, but he refused to acknowledge the results until mass demonstrations and international pressure forced him to step down. In April 2001 he was arrested on charges of corruption and abuse of power.

Indictment of Milosevic

Croatia. In the early 1990s the main spokesman for neofascism in Croatia was Dobroslav Paraga, founder in 1990 of the Croatian Party of Rights (Hrvatska Stranka Prava; HSP). A former seminary student and dissident under the communist regime in Croatia in the 1980s, Paraga believed that Serbia was a mortal danger to Croatian national survival, and he called for the creation of a "Greater Croatia" that would include most areas of Serbia and all of Bosnia and Herzegovina.

Paraga's followers openly endorsed the pro-Nazi Ustaša regime. Reflecting the enthusiasm for Ustaša symbolism that swept Croatia after the outbreak of the Bosnian war in 1991, HSP members often wore caps marked with a U and donned black shirts in imitation of the former Uštasa paramilitary. The HSP's paramilitary wing, the Croatian Defense Association (Hrvatska Obrambeni Savez; HOS), was heavily involved in fighting against Serbia.

Like the SRS in Serbia, the HSP was opposed by a larger ruling party—the Croatian Democratic Union (Hrvatska Demokratska Zajednica; HDZ), founded in 1989 by Franjo Tudjman—that eventually adopted neofascist policies in order to undercut the appeal of its extreme nationalist and neofascist rivals. In 1993 the government launched a largely successful "antifascist" campaign aimed at curbing the influence of HSP supporters in the military, and in the same year Paraga was brought to trial for having allegedly plotted a coup, though he was later acquitted. In 1995 Tudjman's troops undertook extensive ethnic cleansing campaigns in western Slavonia and the historically Serbian region of Krajina, forcing the evacuation of some 150,000 Croatian Serbs to Serbia and Serb-held areas of Bosnia.

Neofascism outside Europe. The largest neofascist movements outside Europe after World War II emerged in Latin America, South Africa, and the Middle East. Juan Perón, who ruled Argentina as the legally elected president in 1946–55 and again in 1973–74, served as a military attaché to Italy in the 1930s and was a great admirer of the Duce.

Perón won the support of poor industrial workers (the *descamisados,* or "shirtless ones") as well as many wealthy businessmen by promoting higher wages and benefits as well as industrial development. He also had the backing of many middle-class nationalists and a large portion of the army officer corps. His charismatic wife, Eva Perón, popularly known as Evita, attracted a cult following for her charitable activities and her storybook rise from "rags to riches." However, owing to inflation, corruption, and Perón's conflicts with the formerly dominant landowning class and the Catholic church, the military eventually turned against him, and he was ousted in a coup in 1955.

After a long exile in Spain, Perón returned to Argentina in 1973 and, in a special election in October of that year, was elected president with his second wife, Isabel Perón, as vice president. Succeeding her husband after his death in 1974, Isabel Perón could not prevent a split between rightist and leftist factions of the Peronist coalition. The economy deteriorated dramatically, and the country was plagued by waves of kidnappings and assassinations of government and business leaders by leftist guerrillas—violence that was soon answered in kind, and on a much larger scale, by the military and secret police. Having lost all popular support, Isabel Perón was overthrown in a military coup in March 1976.

The most significant neofascist group in South Africa after 1945 was the South African Gentile National Socialist Movement (the "Greyshirts"), which changed its name to the White Workers Party in 1949. Although the party did not succeed in creating a mass movement, it did encourage the adoption of policies of white supremacy and apartheid by the dominant National Party of South Africa.

Muammar al-Qaddafi

In the Middle East the regimes of Muammar al-Qaddafi in Libya and Saddam Hussein in Iraq were neofascist in several respects. A charismatic dictator and devout Muslim, Qaddafi came to power in 1969 in a military coup that overthrew King Idris. He advocated what he called "true democracy," characterized by state ownership of key sectors of the economy, strict adherence to Islāmic law, and the mobilization of mass support through "people's congresses," government-controlled labour unions, and other organizations. In Iraq, Hussein's Ba'th movement defended an extremely nationalistic brand of socialism that rejected Western liberalism as well as "materialistic communism." Hussein's regime, which came to power in a coup in 1968, was essentially a personal dictatorship based on an Arab version of the *Führerprinzip.*

In the 1990s a number of "militia" groups in the United States made use of paramilitary uniforms and neo-Nazi symbolism. However, they lacked the popular support necessary to launch a strong political movement or to engage in electoral politics on their own. (Ro.So.)

Nationalism

Nationalism may be defined as a state of mind in which the individual feels that everyone owes his supreme secular loyalty to the nation-state. Nationalism is a modern movement. Throughout history men have been attached to their native soil, to the traditions of their parents, and

to established territorial authorities; but it was not until the end of the 18th century that nationalism began to be a generally recognized sentiment molding public and private life and one of the great, if not the greatest, single determining factors of modern history. Because of its dynamic vitality and its all-pervading character, nationalism is often thought to be very old; sometimes it is mistakenly regarded as a permanent factor in political behaviour. Actually, the American and French revolutions may be regarded as its first powerful manifestations. After penetrating the new countries of Latin America it spread in the early 19th century to central Europe and from there, toward the middle of the century, to eastern and southeastern Europe. At the beginning of the 20th century nationalism flowered in the ancient lands of Asia and Africa. Thus the 19th century has been called the age of nationalism in Europe, while the 20th century has witnessed the rise and struggle of powerful national movements throughout Asia and Africa.

Identification of state and people

Nationalism, translated into world politics, implies the identification of the state or nation with the people—or at least the desirability of determining the extent of the state according to ethnographic principles. In the age of nationalism, but only in the age of nationalism, the principle was generally recognized that each nationality should form a state—its state—and that the state should include all members of that nationality. Formerly states, or territories under one administration, were not delineated by nationality. Men did not give their loyalty to the nation-state but to other, different forms of political organization: the city-state, the feudal fief and its lord, the dynastic state, the religious group, or the sect. The nation-state was nonexistent during the greater part of history, and for a very long time it was not even regarded as an ideal. In the first 15 centuries of the Christian Era, the ideal was the universal world-state, not loyalty to any separate political entity. The Roman Empire had set the great example, which survived not only in the Holy Roman Empire of the Middle Ages but also in the concept of the *res publica christiana* ("Christian republic" or community) and in its later secularized form of a united world civilization.

As political allegiance, before the age of nationalism, was not determined by nationality, so civilization was not thought of as nationally determined. During the Middle Ages civilization was looked upon as determined religiously; for all the different nationalities of Christendom as well as for those of Islām there was but one civilization—Christian or Muslim—and but one language of culture—Latin (or Greek) or Arabic (or Persian). Later, in the periods of the Renaissance and of Classicism, it was the ancient Greek and Roman civilizations that became a universal norm, valid for all peoples and all times. Still later, French civilization was accepted throughout Europe as the valid civilization for educated people of all nationalities. It was only at the end of the 18th century that, for the first time, civilization was considered to be determined by nationality. It was then that the principle was put forward that a man could be educated only in his own mother tongue, not in languages of other civilizations and other times, whether they were classical languages or the literary creations of other peoples who had reached a high degree of civilization.

Cultural nationalism

From the end of the 18th century on, the nationalization of education and public life went hand in hand with the nationalization of states and political loyalties. Poets and scholars began to emphasize cultural nationalism first. They reformed the mother tongue, elevated it to the rank of a literary language, and delved deep into the national past. Thus they prepared the foundations for the political claims for national statehood soon to be raised by the people in whom they had kindled the spirit.

Before the 18th century there had been evidences of national feeling among certain groups at certain periods, especially in times of stress and conflict. The rise of national feeling to major political importance was encouraged by a number of complex developments: the creation of large, centralized states ruled by absolute monarchs who destroyed the old feudal allegiances; the secularization of life and of education, which fostered the vernacular languages and weakened the ties of church and sect; the growth of commerce, which demanded larger territorial units to allow scope for the dynamic spirit of the rising middle classes and their capitalistic enterprise. This large, unified territorial state, with its political and economic centralization, became imbued in the 18th century with a new spirit—an emotional fervour similar to that of religious movements in earlier periods. Under the influence of the new theories of the sovereignty of the people and the rights of man, the people replaced the king as the centre of the nation. No longer was the king the nation or the state; the state had become the people's state, a national state, a fatherland. State became identified with nation, as civilization became identified with national civilization.

That development ran counter to the conceptions that had dominated political thought for the preceding 2,000 years. Hitherto man had commonly stressed the general and the universal and had regarded unity as the desirable goal. Nationalism stressed the particular and parochial, the differences, and the national individualities. Those tendencies became more pronounced as nationalism developed. Its less attractive characteristics were not at first apparent. In the 17th and 18th centuries the common standards of Western civilization, the regard for the universally human, the faith in reason (one and the same everywhere) as well as in common sense, the survival of Christian and Stoic traditions—all of these were still too strong to allow nationalism to develop fully and to disrupt society. Thus nationalism in its beginning was thought to be compatible with cosmopolitan convictions and with a general love of mankind, especially in western Europe and North America.

EUROPEAN NATIONALISM

The first full manifestation of modern nationalism occurred in 17th-century England, in the Puritan revolution. England had become the leading nation in scientific spirit, in commercial enterprise, in political thought and activity. Swelled by an immense confidence in the new age, the English people felt upon their shoulders the mission of history, a sense that they were at a great turning point from which a new true reformation and a new liberty would start. In the English revolution an optimistic humanism merged with Calvinist ethics; the influence of the Old Testament gave form to the new nationalism by identifying the English people with ancient Israel.

English Puritanism and nationalism

The new message, carried by the new people not only for England but for all mankind, was expressed in the writings of John Milton, in whose famous vision the idea of liberty was seen spreading from Britain, "celebrated for endless ages as a soil most genial to the growth of liberty" to all the corners of the earth.

> Surrounded by congregated multitudes, I now imagine that . . . I behold the nations of the earth recovering that liberty which they so long had lost; and that the people of this island are . . . disseminating the blessings of civilization and freedom among cities, kingdoms and nations.

English nationalism then was thus much nearer to its religious matrix than later nationalisms that rose after secularization had made greater progress. The nationalism of the 18th century shared with it, however, its enthusiasm for liberty, its humanitarian character, its emphasis upon the individual and his rights and upon the human community as above all national divisions. The rise of English nationalism coincided with the rise of the English trading middle classes. It found its final expression in John Locke's political philosophy, and it was in that form that it influenced American and French nationalism in the following century.

American nationalism was a typical product of the 18th century. British settlers in North America were influenced partly by the traditions of the Puritan revolution and the ideas of Locke and partly by the new rational interpretation given to English liberty by contemporary French philosophers. American settlers became a nation engaged in a fight for liberty and individual rights. They based that fight on current political thought, especially as expressed by Thomas Jefferson and Thomas Paine. It was a liberal and humanitarian nationalism that regarded America as in the vanguard of mankind on its march to

greater liberty, equality, and happiness for all. The ideas of the 18th century found their first political realization in the Declaration of Independence and in the birth of the American nation. Their deep influence was felt in the French Revolution.

Jean-Jacques Rousseau had prepared the soil for the growth of French nationalism by his stress on popular sovereignty and the general cooperation of all in forming the national will, and also by his regard for the common people as the true depository of civilization.

The nationalism of the French Revolution was more than that: it was the triumphant expression of a rational faith in common humanity and liberal progress. The famous slogan "liberty, equality, fraternity" and the Declaration of the Rights of Man and of the Citizen were thought valid not only for the French people but for all peoples. Individual liberty, human equality, fraternity of all peoples: these were the common cornerstones of all liberal and democratic nationalism. Under their inspiration new rituals were developed that partly took the place of the old religious feast days, rites, and ceremonies: festivals and flags, music and poetry, national holidays and patriotic sermons. In the most varied forms, nationalism permeated all manifestations of life. As in America, the rise of French nationalism produced a new phenomenon in the art of warfare: the nation in arms. In America and in France, citizen armies, untrained but filled with a new fervour, proved superior to highly trained professional armies that fought without the incentive of nationalism. The revolutionary French nationalism stressed free individual decision in the formation of nations. Nations were constituted by an act of self-determination of their members. The plebiscite became the instrument whereby the will of the nation was expressed. In America as well as in revolutionary France, nationalism meant the adherence to a universal progressive idea, looking toward a common future of freedom and equality, not toward a past characterized by authoritarianism and inequality.

Napoleon's armies spread the spirit of nationalism throughout Europe and even into the Near East, while at the same time, across the Atlantic, it aroused the Latin Americans. But Napoleon's yoke of conquest turned the nationalism of the Europeans against France. In Germany the struggle was led by writers and intellectuals, who rejected all the principles upon which the American and the French revolutions had been based as well as the liberal and humanitarian aspects of nationalism.

German nationalism began to stress instinct against reason; the power of historical tradition against rational attempts at progress and a more just order; the historical differences between nations rather than their common aspirations. The French Revolution, liberalism, and equality were regarded as a brief aberration, against which the eternal foundations of societal order would prevail.

That German interpretation was shown to be false by the developments of the 19th century. Liberal nationalism reasserted itself and affected more and more people: the rising middle class and the new proletariat. The revolutionary wave of 1848, the year of "the spring of the peoples," seemed to realize the hopes of nationalists such as Giuseppe Mazzini, who had devoted his life to the unification of the Italian nation by democratic means and to the brotherhood of all free nations. Though his immediate hopes were disappointed, the 12 years from 1859 to 1871 brought the unification of Italy and Romania, both with the help of Napoleon III, and of Germany; at the same time the 1860s saw great progress in liberalism, even in Russia and Spain. The victorious trend of liberal nationalism, however, was reversed in Germany by Bismarck. He unified Germany on a conservative and authoritarian basis and defeated German liberalism. The German annexation of Alsace-Lorraine against the will of the inhabitants was contrary to the idea of nationalism as based upon the free will of man. The people of Alsace-Lorraine were held to be German by objective factors, by race, independent of their will or of their allegiance to any nationality of their choice.

The 1848 revolutionary wave

In the second half of the 19th century, nationalism disintegrated the supranational states of the Habsburgs and the Ottoman sultans, both of which were based upon prenational loyalties. In Russia, the penetration of nationalism produced two opposing schools of thought. Some nationalists proposed a westernized Russia, associated with the progressive, liberal forces of the rest of Europe. Others stressed the distinctive character of Russia and Russianism, its independent and different destiny based upon its autocratic and orthodox past. These Slavophiles, similar to and influenced by German romantic thinkers, saw Russia as a future saviour of a West undermined by liberalism and the heritage of the American and French revolutions.

One of the consequences of World War I was the triumph of nationalism in central and eastern Europe. From the ruins of the Habsburg and Romanov empires emerged the new nation-states of Austria, Hungary, Czechoslovakia, Poland, Yugoslavia and Romania. Those states in turn, however, were to be strained and ravaged by their own internal nationality conflicts and by nationalistic disputes over territory with their neighbours.

Russian nationalism was in part suppressed after Lenin's victory in 1917, when the Bolsheviks took over the old empire of the tsars. But the Bolsheviks also claimed the leadership of the world Communist movement, which was to become an instrument of the national policies of the Russians. During World War II Stalin appealed to nationalism and patriotism in rallying the Russians against foreign invaders. After the war he found nationalism one of the strongest obstacles to the expansion of Soviet power in eastern Europe. National communism, as it was called, became a divisive force in the Soviet bloc. In 1948 Tito, the Communist leader of Yugoslavia, was denounced by Moscow as a nationalist and a renegade; nationalism was a strong factor in the rebellious movements in Poland and Hungary in the fall of 1956; and subsequently its influence was also felt in Romania and Czechoslovakia and again in Poland in 1980.

ASIAN AND AFRICAN NATIONALISM

Nationalism began to appear in Asia and Africa after World War I. It produced such leaders as Kemal Atatürk in Turkey, Sa'd Pasha Zaghūl in Egypt, Ibn Sa'ūd in the Arabian peninsula, Mahatma Gandhi in India, and Sun Yat-sen in China. Atatürk succeeded in replacing the medieval structure of the Islāmic monarchy with a revitalized and modernized secular republic in 1923. Demands for Arab unity were frustrated in Africa and Asia by British imperialism and in Africa by French imperialism. Yet Britain may have shown a gift for accommodation with the new forces by helping to create an independent Egypt (1922; completely, 1936) and Iraq (1932) and displayed a similar spirit in India, where the Indian National Congress, founded in 1885 to promote a liberal nationalism inspired by the British model, became more radical after 1918. Japan, influenced by Germany, used modern industrial techniques in the service of a more authoritarian nationalism.

The new nations

The progress of nationalism in Asia and Africa is reflected in the histories of the League of Nations after World War I and of the United Nations after World War II. The Treaty of Versailles, which provided for the constitution of the League of Nations, also reduced the empires of the defeated Central Powers, mainly Germany and Turkey. The league distributed Germany's African colonies as mandates to Great Britain, France, Belgium, and South Africa, and its Pacific possessions to Japan, Australia, and New Zealand under various classifications according to their expectations of achieving independence. Among the League's original members, there were only five Asian countries (China, India, Japan, Thailand, and Iran) and two African countries (Liberia and South Africa), and it added only three Asian countries (Afghanistan, Iraq, and Turkey) and two African countries (Egypt and Ethiopia) before it was dissolved in 1946. Of the mandated territories under the League's control, only Iraq, Lebanon, and Syria achieved independence during its lifetime.

Of the original 51 members of the United Nations in 1945, eight were Asian (China, India, Iraq, Iran, Lebanon, Saudi Arabia, Syria, and Turkey) and four were African (the same as in the League). By 1980, 35 years after

its founding, the United Nations had added more than 100 member nations, most of them Asian and African. Whereas Asian and African nations had never totalled even one-third of the membership in the League, they came to represent more than one-half of the membership of the United Nations. Of these new nations, several had been created, entirely or in part, from mandated territories.

After World War II, India, Pakistan, Ceylon (Sri Lanka), Burma and Malaya (Malaysia) in Asia, and Ghana in Africa achieved independence peacefully from the British Commonwealth, as did the Philippines from the United States. Other territories had to fight hard for their independence in bitter colonial wars, as in French Indochina (Vietnam, Laos, Cambodia) and French North Africa (Tunisia, Algeria). Communism recruited supporters from within the ranks of the new nationalist movements in Asia and Africa, first by helping them in their struggles against Western capitalist powers and later, after independence was achieved by competing with Western capitalism in extending financial and technical aid. Chinese nationalism under Chiang Kai-shek during World War II was diminished with the takeover of the Chinese Communists. But Chinese Communism soon began to drift away from supranational Communism, as the European Communist countries had earlier. During the 1960s Chinese Communism turned further and further inward and its influence on new Asian and African nations waned. After the fall of Communism in the Soviet Union and eastern Europe in 1990–91, China embraced and expanded market-oriented economic reforms and undertook to modernize its military in order to play a more assertive role in world affairs. Elsewhere in Asia, the international financial crisis of the late 1990s prompted many political leaders to criticize the International Monetary Fund (IMF) and other world bodies for infringing on their countries' national sovereignty.

Political and religious differences

Ambitions among new Asian and African nations clashed. The complex politics of the United Nations illustrated the problems of the new nationalism. The struggle with Dutch colonialism that brought the establishment of Indonesia continued with the UN mediation of the dispute over West Irian (Irian Jaya). In the Suez crisis of 1956, UN forces intervened between those of Egypt and Israel. Continuing troubles in the Middle East, beginning with the establishment of Israel and including inter-Arab state disputes brought on by the establishment of the United Arab Republic, concerned the UN. Other crises involving the UN included: the India-Pakistan dispute over Jammu and Kashmir; the Korean partition and subsequent war; the four-year intervention in the Congo; the struggle of Greece and Turkey over newly independent Cyprus; Indonesian and Philippine objection to the inclusion of Sarawak and Sabah (North Borneo) in newly formed Malaysia; and the destruction of Iraq's medium-range missiles and chemical-, biological-, and nuclear-weapons research facilities after the Persian Gulf War (1990–91).

Many new nations, all sharing the same pride in independence, faced difficulties. As a result of inadequate preparation for self-rule, the first five years of independence in the Congo passed with no semblance of a stable government. The problem of widely different peoples and languages was exemplified in Nigeria, where an uncounted population included an uncounted number of tribes (at least 150, with three major divisions) that used an uncounted number of languages (more than 100 language and dialect clusters). The question of whether the predominantly Muslim state of Jammu and Kashmir should go with Muslim Pakistan or Hindu India lasted for more than 20 years after the India Independence Act became effective in 1949. Desperate economic competition caused trouble, as in Israel where the much-needed waters of the Jordan River kept it in constant dispute with its water-hungry Arab neighbours. (H.K./Ed.)

Liberalism

Liberalism is the political doctrine that takes the abuse of power, and thus the freedom of the individual, as the central problem of government. For liberals, power is most importantly abused by governments, but it may also be abused by the wealthy, by monarchs, aristocrats, and others with inherited authority and privileges; and indeed by any group that has the means and the inclination to act oppressively.

Two meanings of liberalism

Historically, liberalism has come to mean two rather different things. The doctrine originated as a defensive reaction to the horrors of the wars of religion of the 16th century and then divided into two strands, the first a narrowly political doctrine emphasizing the importance of limited government, the other a philosophy of life emphasizing individual autonomy, imagination, and self-development. In addition, contemporary liberalism has come to represent different things to Americans and Europeans: In the United States it is associated with the welfare-state policies of Democratic President Franklin D. Roosevelt, whereas in Europe liberals are more commonly conservative in their political and economic outlook.

Liberalism derives from two related features of Western culture. The first is the West's preoccupation with individuality, as compared to the emphasis in other civilizations on status, caste, and tradition. Throughout much of history, the individual has been submerged in his clan, tribe, people, or kingdom. Liberalism is the culmination of developments in Western society that produced a sense of the importance of human individuality, a liberation of the individual from complete subservience to the group, and a relaxation of the tight hold of custom, law, and authority. The emancipation of the individual can be understood as a unique achievement of Western culture, perhaps its very hallmark.

Liberalism also derives from the practice of adversariality in European political and economic life, a process in which institutionalized competition—such as the competition between different political parties in electoral contests, between prosecution and defense in judicial procedures, or between different producers in a free-market economy—is used to generate a dynamic social order. Adversarial systems have always been precarious, however, and it took a long time for the belief in adversariality to emerge from the more traditional view, traceable at least to Plato, that the state should be an organic structure in which the different social classes cooperate by performing distinct yet complementary roles. The belief that competition is an essential part of a political system and that good government requires a vigorous opposition was still considered strange in most European countries in the early 19th century.

Underlying the liberal belief in adversariality is the conviction that human beings are essentially rational creatures capable of settling their political disputes through dialogue and compromise. This aspect of liberalism became particularly prominent in 20th-century projects aimed at eliminating war and resolving disagreements between states through organizations such as the United Nations.

It is evident that liberalism has a close relationship with democracy, but not too much should be made of this association. At the centre of democratic doctrine is the belief that governments derive their authority from popular election; liberalism, on the other hand, is primarily concerned with the scope of governmental activity. Liberals often have been wary of democracy because of fears that it might generate a tyranny by the majority. One might briskly say, therefore, that democracy looks after majorities and liberalism after minorities.

Like other political doctrines, liberalism is highly sensitive to time and circumstance. Each nation's liberalism is different, and it changes in each generation. The historical development of liberalism over recent centuries has been a movement from mistrust of the state's sovereignty on the ground that power tends to be misused, to a willingness to use the power of government to correct inequities in the distribution of the wealth resulting from the market economy. The expansion of government power and responsibility sought by liberals in the 20th century was clearly opposed to the contraction of government advocated by liberals a century earlier. In the 19th century liberals were generally hospitable to the business community, only to become hostile to its interests and ambitions for much of the 20th century. In each case, however, the liberals' inspiration was the same: a hostility to concentrations of power that threaten the freedom of the individual and prevent him from realizing his potential, along with a willingness

to reexamine and reform social institutions in the light of new needs. This willingness is tempered by an aversion to sudden, cataclysmic change, which is what sets off the liberal from the radical. It is this very eagerness to consider and encourage useful change, however, that distinguishes the liberal from the conservative.

CLASSICAL LIBERALISM

Political foundations. Although liberal ideas were not noticeable in European politics until the early 16th century, liberalism has a considerable "prehistory" reaching back to the Middle Ages and even earlier. In the Middle Ages the rights and responsibilities of the individual were determined by his place in a stratified hierarchy that placed great stress upon acquiescence and conformity. Under the impact of the slow commercialization and urbanization of Europe in the later Middle Ages, the intellectual ferment of the Renaissance, and the spread of Protestantism in the 16th century, the old feudal stratification of society gradually began to dissolve, leading to a fear of instability so powerful that monarchical absolutism was viewed as the obvious solution to civil dissension. By the end of the 16th century, the authority of the papacy had been broken in most of northern Europe, and each ruler tried to consolidate the unity of his realm by enforcing religious uniformity. These efforts culminated in the Thirty Years' War, which did immense damage to much of Europe. Where no creed succeeded in wholly extirpating its enemies, toleration was gradually accepted as the lesser of two evils; in some countries where one creed triumphed, it was accepted that too minute a concern with citizens' beliefs was inimical to prosperity and good order.

The ambitions of national rulers and the requirements of expanding industry and commerce led gradually to the adoption of economic policies based on mercantilism, a school of thought that promoted government regulation of a nation's economy to increase state wealth and power at the expense of rival nations. However, as such intervention increasingly served established interests and inhibited enterprise, it was challenged by members of the newly emerging middle class. This challenge was a significant factor in the great revolutions that rocked England and France in the 17th and 18th centuries—most notably the English Civil Wars from 1642 to 1651, the English Revolution of 1688 (the "Glorious Revolution"), the United States War of Independence from 1775 to 1783, and the French Revolution of 1789. Classical liberalism as an articulated creed is a product of those great collisions.

In the English Civil Wars, the absolutist king Charles I was defeated by the forces of Parliament and eventually executed. The Revolution of 1688 resulted in the abdication and exile of James II and the establishment of a complex form of balanced government in which power was divided between the king, his ministers, and Parliament. In time this system would become a model for liberal political movements in other countries. The political ideas that helped to inspire these revolts were given formal expression in the work of the English philosophers Thomas Hobbes and John Locke. In *Leviathan* (1651), Hobbes argued that the absolute power of the sovereign was ultimately justified by the consent of the governed, who agreed—in a hypothetical "social contract"—to obey the sovereign's will in all matters in exchange for a guarantee of peace and security. Locke, who also held a social-contract theory of government, argued that it was the role of the sovereign to protect the person and property of individuals and to guarantee their natural rights to freedom of thought, speech, and worship. Significantly, Locke thought that revolution was legitimate in cases where the sovereign failed to fulfill these obligations, and it was long assumed that *Two Treatises of Government* (1690), his major work in political theory, was written precisely in order to justify the revolution of two years before.

Thomas Hobbes and John Locke

By the time Locke had written his *Treatises,* politics in England had become a contest between two loosely related parties whose members were known as Whigs and Tories. Locke was a notable Whig, and it is conventional to view liberalism as derived from the attitudes of Whig aristocrats, who were often linked with commercial interests and who had an entrenched suspicion of the power of the monarchy. The Whigs dominated English politics from the death of Queen Anne in 1714 to the accession of King George III in 1760.

Economic foundations. If the political foundations of liberalism were laid in Great Britain, so too were its economic foundations. By the 18th century British monarchs were constrained by Parliament from pursuing the schemes of national aggrandizement favoured by most rulers on the Continent. These rulers fought for military supremacy, which required a strong economic base. Because the prevailing mercantilist theory understood international trade as a zero-sum game—in which gain for one nation meant loss for another—national governments intervened to determine prices, protect their industries from foreign competition, and avoid the sharing of economic information.

These practices were challenged by the Scottish economist and philosopher Adam Smith, who argued in *The Wealth of Nations* (1776) that free trade would benefit all parties. According to this view, if individuals are left free to pursue their self-interest in an exchange economy based upon a division of labour, the welfare of the group as a whole necessarily will be enhanced. Smith described a self-adjusting market mechanism whose one propelling force is the selfishness of the individual. The self-seeking individual becomes harnessed to the public good because in an exchange economy he must serve others in order to serve himself. But it is only in a genuinely free market, according to Smith, that this positive consequence is possible; any other arrangement, whether state control or monopoly, must lead to regimentation, exploitation, and economic stagnation.

The Wealth of Nations

Every economic system must determine not only what goods will be produced but also how those goods are to be apportioned, or distributed. In a free-market economy both of these tasks are accomplished through the price mechanism. The theoretically free choices of individual buyers and sellers determine how the resources of society—labour, goods, and capital—shall be employed. These choices manifest themselves in bids and offers that together determine a commodity's price. Theoretically, when the demand for a commodity is great, prices rise, making it profitable for producers to increase the supply; as supply approximates demand, prices tend to fall until producers divert productive resources to other uses. In this way the system achieves the closest possible match between what is desired and what is produced. Moreover, in the distribution of the wealth thereby produced, the system is said to assure a reward in proportion to merit. The assumption is that, in a freely competitive economy in which no one is barred from engaging in economic activity, the income received from such activity is a fair measure of its value to society.

Presupposed in the foregoing account is a conception of human beings as economic animals rationally and self-interestedly engaged in minimizing costs and maximizing gains. If, as Enlightenment liberals assumed, human beings rarely fail to act rationally in any of their activities, it follows that people would meticulously balance benefits against costs in the marketplace. Since each human being knows his own interests better than anyone else does, his interests could only be hindered, and never enhanced, by government interference in his economic activities.

In concrete terms, classical liberal economists called for several major changes in the sphere of British and European economic organization. The first was the abolition of the host of feudal and mercantilist restrictions on nations' manufacturing and internal commerce. The second was an end to the tariffs and restrictions that governments imposed on foreign imports to protect domestic producers. In rejecting the government's regulation of trade, classical economics was based firmly on a belief in the superiority of a self-regulating market. Quite apart from the cogency of their arguments, the views of Smith and his 19th-century English successors, the economist David Ricardo and the philosopher and economist John Stuart Mill, became increasingly convincing as Britain's Industrial Revolution generated enormous new wealth and made that country into the "workshop of the world." Several generations of Europeans were eventually persuaded that policies of free trade would make everyone prosperous.

Principle of limited government

Inspired by the need to remove the state from destructive interference with economic life, the guiding political principle of classical liberalism became an undeviating insistence on limiting the power of government. The English Utilitarian philosopher Jeremy Bentham cogently summarized this view in his sole advice to the state: "Be quiet"; and the American statesman Thomas Jefferson asserted that that government is best that governs least. Classical liberals freely acknowledged that government must provide education, sanitation, law enforcement, postal delivery, and other public services that were beyond the capacity of any private agency. But liberals generally believed that, apart from these functions, government must not do for the individual what he is able to do for himself.

Liberalism and Utilitarianism. In the late 18th and early 19th centuries, Bentham, the Utilitarian philosopher James Mill, and James's son John Stuart Mill applied classical economic principles to the political sphere. Invoking the concept of utility, they argued that the object of all legislation should be "the greatest happiness of the greatest number." In evaluating what kind of government could best attain this objective, the Utilitarians generally supported representative democracy, asserting that it was the best way to make the interest of government coincide with the general interest. Taking their cue from the notion of a free-market economy, the Utilitarians called for a political system that would guarantee its citizens the maximum degree of individual freedom of choice and action consistent with efficient government and the preservation of social harmony. They advocated expanded education, enlarged suffrage, and periodic elections to ensure government's accountability to the governed. They also developed a doctrine of individual rights—including the rights to freedom of religion, freedom of speech, freedom of the press, and freedom of assembly—that lies at the heart of modern democracy. These rights received their classic advocacy in John Stuart Mill's essay *On Liberty* (1859), which argues on Utilitarian grounds that the state may regulate individual behaviour only in cases where the interests of others would be perceptibly harmed. Today, this work is justly celebrated as one of the great testimonials to civil liberties and as a classic of liberal thought.

The Utilitarians thus succeeded in broadening the philosophical foundations of political liberalism while also providing a program of specific reformist goals for liberals to pursue. Their overall political philosophy was perhaps best stated in James Mill's article "Government," which was written for the supplement (1815–24) to the fourth through sixth editions of the *Encyclopædia Britannica.*

Liberalism and democracy. Politically, liberalism ultimately aspired to a system of government based on majority rule—*i.e.,* one in which government executed the expressed will of a majority of the electorate. The chief institutional devices for attaining this goal were the periodic election of legislators by popular vote and the election of a chief executive by popular vote or by a legislative assembly.

But, in answering the crucial question of who is to be the electorate, classical liberalism fell victim to ambivalence, torn between the great emancipating tendencies generated by the revolutions with which it was associated and middle-class fears that a wide or universal franchise would undermine private property. Benjamin Franklin spoke for the Whig liberalism of the founding fathers of the United States when he stated, "As to those who have not landed property the allowing them to vote is an impropriety." John Adams, in his famous *Defense of the Constitutions of Government of the United States of America* (1787), was more explicit, finding that, if the majority were to control all branches of government, "Debts would be abolished first; taxes laid heavy on the rich, and not at all on others; and at last a downright equal division of everything be demanded and voted." French statesmen such as François Guizot and Adolphe Thiers expressed similar sentiments well into the 19th century.

Most 18th- and 19th-century liberal spokesmen thus feared popular sovereignty, and for a long time suffrage was limited to property owners. In Britain even the important Reform Act of 1867 did not completely abolish property qualifications for the right to vote. In France, although the Revolution of 1789 proclaimed the ideal of universal manhood suffrage and the Revolution of 1830 reaffirmed it, there were no more than 200,000 qualified voters in a population of about 30 million during the reign of Louis-Philippe, the "citizen king" (1830–48). In the United States, Thomas Jefferson's brave language in the Declaration of Independence notwithstanding, it was not until 1860 that universal white male suffrage prevailed. In most of Europe universal male suffrage remained a remote ideal until late in the 19th century.

Despite the misgivings of the propertied classes, a slow but steady expansion of the franchise prevailed throughout Europe in the 19th century. But the principle of majority rule also had to be reconciled with the liberal requirement that the power of the majority be a limited one. The problem was to accomplish this in a manner consistent with the democratic ideal. Given that hereditary elites were discredited, how could the power of the majority be checked without giving disproportionate power to property owners or to some other "natural" elite?

Separation of powers. The liberal solution to the problem of limiting the powers of a democratic majority rested on various devices. The first was the separation of powers—*i.e.,* the distribution of power between functionally differentiated agencies of government such as the legislature, the executive, and the judiciary. This arrangement, and the system of checks and balances by which it was accomplished, was given its classic embodiment in the Constitution of the United States and its political justification in *The Federalist* (1788), by Alexander Hamilton, James Madison, and John Jay. Of course, such a separation of powers also could have been achieved through a "mixed constitution"—*i.e.,* one by which a monarch, a hereditary chamber, and an elected assembly share power with some appropriate differentiation of function. This was in fact the system of government in Great Britain at the time of the American Revolution. But despotic kings and functionless aristocrats (more functionless in France than in England) thwarted the interests and ambitions of the middle class, which turned, therefore, to the principle of majoritarianism.

The Federalist

Periodic elections. The second part of the solution lay in using staggered periodic elections to make the decisions of any given majority subject to the concurrence of other majorities distributed over time. In the United States, for example, presidents are elected every four years and members of the House of Representatives every two years, and one-third of the Senate is elected every two years for terms of six years. Therefore, the majority that elects a president every four years or a House of Representatives every two years is different from the majority that elects one-third of the Senate two years earlier and the majority that elects another one-third two years later. These bodies, in turn, are "checked" by the Constitution, which was approved and amended by earlier majorities. In Britain an act of Parliament immediately becomes part of the unwritten constitution; however, before acting on a highly controversial issue, Parliament must seek a mandate from the people, which represents a majority other than the one that elected it. Thus, in a constitutional democracy, the power of a current majority is checked by the verdicts of majorities that precede and follow it.

Rights. The third part of the solution was related to liberalism's basic commitment to the autonomy and integrity of the individual, which the limitation of power is, after all, intended to preserve. In the liberal understanding, the individual is not only a citizen who shares a social compact with his fellows but also a person with rights upon which the state may not encroach if majoritarianism is to be meaningful. A majority verdict can come about only if individuals are free to some extent to exchange their views. This involves, beyond the right to speak and write freely, the freedom to associate and organize and, above all, the freedom from fear of reprisal. But the individual also has rights apart from his role as citizen. These rights secure his personal safety and hence his protection from arbitrary arrest and punishment. Beyond these rights are those that preserve large areas of privacy. In a liberal democracy there are affairs that do not concern the state. Such affairs may range from the practice of religion, to the creation of art,

to the raising of children. For liberals of the 18th and 19th centuries they included, above all, most of the activities through which individuals engage in production and trade.

The Declaration of Independence

Eloquent and persuasive declarations affirming such rights were embodied in the English Bill of Rights of 1689, the United States Declaration of Independence and Constitution (1776 and 1788, respectively), the French Declaration of the Rights of Man and of the Citizen of 1789, and the basic documents of nations throughout the world that later used these declarations as their models. Freedom thereby became more than the right to make a fractional contribution in an intermittent mandate to government; it designated the right of people to live their own lives.

LIBERALISM IN THE 19TH CENTURY

As an ideology and in practice liberalism became the preeminent reform movement in Europe during the 19th century. Its fortunes, however, differed with the historical conditions in each country—the strength of the crown, the élan of the aristocracy, the pace of industrialization, and the circumstances of national unification. The national character of a liberal movement could even be affected by religion. Liberalism in Roman Catholic countries such as France, Italy, and Spain, for example, tended to acquire anticlerical overtones, and liberals in those countries tended to favour legislation restricting the civil authority and political power of the Catholic clergy.

In Great Britain the Whigs had evolved by the mid-19th century into the Liberal Party, whose reformist programs became the model for liberal political parties throughout Europe. Liberals propelled the long campaign that abolished Britain's slave trade in 1807 and slavery itself throughout the British dominions in 1833. The liberal project of broadening the franchise in Britain bore fruit in the Reform Bills of 1832, 1867, and 1884–85. The sweeping reforms achieved by Liberal Party governments led by William Gladstone for 14 years between 1868 and 1894 marked the apex of British liberalism.

Liberalism in Europe often lacked the fortuitous combination of broad popular support and a powerful liberal party that it had in Britain. In France the Revolutionary and Napoleonic governments pursued liberal goals in their abolition of feudal privileges and their modernization of the decrepit institutions inherited from the ancien régime. After the Bourbon restoration in 1815, however, French liberals had the decades-long task of securing constitutional liberties and enlarging popular participation in government in the face of a reestablished monarchy, goals not substantially achieved until the formation of the Third Republic in 1871.

Throughout Europe and in the Western Hemisphere, liberalism inspired nationalistic aspirations to the creation of unified, independent, constitutional states with their own parliaments and the rule of law. The most dramatic exponents of this liberal assault against authoritarian rule were the founding fathers in the United States, the statesman and revolutionary Simón Bolívar in South America, the leaders of the Risorgimento in Italy, and the nationalist reformer Lajos Kossuth in Hungary. But the failure of the Revolutions of 1848 highlighted the comparative weakness of liberalism on the Continent. Liberals' inability to unify the German states in the mid-19th century was attributable in large part to the dominant role of a militarized Prussia and the reactionary influence of Austria. The liberal-inspired unification of Italy was delayed until the 1860s by the armies of Austria and of Napoleon III of France and by the opposition of the Vatican.

The United States presented a quite different situation, because there was neither a monarchy, an aristocracy, nor an established church to which liberalism could react. Indeed, liberalism was so well established in the United States' constitutional structure, its political culture, and its jurisprudence that there was no distinct role for a liberal party to play, at least not until the 20th century.

Liberalism ended up transforming Europe in the 19th century. The forces of industrialization and modernization, for which classical liberalism was the rationalization, wrought great changes. The feudal system was destroyed, a functionless aristocracy was deprived of its privileges, and monarchs were challenged and curbed. Capitalism replaced the static economies of the Middle Ages, and the middle class was left free to employ its energies expanding the means of production and vastly increasing the wealth of society. Representative government came into its own, and, as liberals set about limiting the power of the monarchy, they converted the ideal of constitutional government into a reality.

MODERN LIBERALISM

Problems of market economies. By the end of the 19th century, some unforeseen but serious consequences of the Industrial Revolution in Europe and North America had produced a deepening disenchantment with the principal economic basis of classical liberalism—the ideal of a market economy. The main problem was that the profit system had concentrated vast wealth in the hands of a relatively small number of industrialists and financiers, with several decisively adverse consequences. First, great masses of people failed to benefit from the wealth flowing from factories and lived in poverty in the slums of dreary cities. Second, because the vastly expanded system of production created many goods and services that people often could not afford to buy, markets became glutted and the system periodically came to a near halt in periods of stagnation that came to be called "depressions." Finally, those who owned or managed the means of production had acquired vast economic power that they used to influence and control government, to manipulate an inchoate electorate, to limit competition, and to obstruct substantive social reform. In short, some of the same forces that had once released the productive energies of Western society now restrained them; some of the very energies that had demolished the power of despots now nourished a new despotism. Such, at any rate, was the verdict of 20th-century liberals.

Despotism of the wealthy

The modern liberal program. In attempting to correct the problems that accompanied industrialization, classical liberalism underwent several major modifications. Most notably, its traditional emphasis on minimizing the role and power of government was reversed. By the early 20th century, liberals instead had begun looking to government to minimize economic inequalities and prevent the exploitation of labour and the abuses of monopolies by redistributing wealth and regulating private industry. The result, which may be termed modern liberalism, has at times been difficult to distinguish from the social democracy movement that arose among the European working classes in the late 19th century. Nevertheless, the outlines of modern liberalism are fairly discernible.

Limited intervention in the market. Because they appreciated the real achievements of the free-market system, modern liberals did not seek its abolition but rather its modification and control. They saw no reason for a fixed line eternally dividing the private and public sectors of the economy; the division, they contended, must be made by reference to what works. The spectre of regimentation in centrally planned economies and the dangers of bureaucracy even in mixed economies deterred them from jettisoning the market and substituting a putatively omnicompetent state. On the other hand—and this is a basic difference between classical and modern liberalism—most liberals came to recognize that the operation of the market needed to be supplemented and corrected in substantive ways. Liberals asserted that the rewards dispensed by the market were too crude a measure of the contribution most people made to society and that the market ignored the needs of those who lacked opportunity or who were economically exploited. They contended that the enormous social costs incurred in production were not reflected in market prices and that resources were used wastefully. Not least, liberals perceived that the market biased the allocation of human and physical resources toward the satisfaction of consumer appetites—*e.g.*, for automobiles, home appliances, or fashionable clothing—while basic needs—*e.g.*, for schools, housing, public transit, and sewage treatment—went unmet. Finally, although liberals believed that prices, wages, and profits should continue to be subject to negotiation among the interested parties and responsive to conventional market pressures, they insisted that price-wage-profit decisions affecting the economy as a whole must be reconciled with public policy.

Greater equality of wealth and income. To achieve a more just distribution of wealth and income, liberals relied on two major strategies. First, they promoted the organization of workers into trade unions in order to improve the workers' power to bargain with employers. Such a redistribution of power had political as well as economic consequences, making possible a multiparty system in which at least one party was responsive to the interests of wage earners.

Introduction of social services

Second, enlisting the political support of the economically deprived, liberals introduced a variety of government-funded social services. Beginning with free public education and workmen's accident insurance, these services later came to include programs of old-age, unemployment, and health insurance; minimum-wage laws; and support for the physically and mentally handicapped. Meeting these objectives required a redistribution of wealth that was achieved by graduated income and inheritance taxes, which affected the wealthy more than they did the poor. Social-welfare measures were first undertaken in Germany in the late 19th century and were soon adopted by other countries of western Europe. In the United States such measures were not adopted at the federal level until passage of the Social Security Act of 1935.

World War I and the Great Depression. The further development of liberalism in Europe was brutally interrupted in 1914–18 by World War I. The war overturned four of Europe's great imperial dynasties—Germany, Austria-Hungary, Russia, and Ottoman Turkey—and thus at first appeared to give added impetus to liberal democracy. Europe was reshaped by the Treaty of Versailles on the principle of national self-determination, which in practice meant the breakup of the German, Austro-Hungarian, and Ottoman empires into nationally homogeneous states. The League of Nations was created in the hope that negotiation would replace war as a means of settling international disputes.

But the trauma of the war had created widespread disillusionment about the entire liberal view of progress toward a more humane world. The harsh peace terms imposed by the victorious Allies, together with the misery created by the Great Depression beginning in 1929, enfeebled Germany's newly established republic and set the stage for the Nazi seizure of power in 1933. In Italy, meanwhile, dissatisfaction with the peace settlement led directly to the Fascist takeover in 1922. Liberalism was also threatened by Soviet Communism, which seemed to many to have inherited the hopes for progress earlier associated with liberalism itself.

If liberalism came under political attack in the interwar period, the future of the market economy was called into question by the Great Depression. The boom-and-bust character of the business cycle had long been a major defect of market economies, but the Great Depression, with its seemingly endless downturn in business activity and its soaring levels of unemployment, confounded classical economists and produced real pessimism about the viability of capitalism.

The wrenching hardships inflicted by the Great Depression eventually convinced Western governments that complex modern societies needed some measure of rational economic planning. The New Deal (1933–39), the domestic program undertaken by U.S. President Franklin D. Roosevelt to lift the United States out of the Great Depression, typified modern liberalism in its vast expansion of the scope of governmental activities and its increased regulation of business. Among the measures that New Deal legislation provided were emergency assistance and temporary jobs to the unemployed, restrictions on banking and financial industries, more power for trade unions to organize and bargain with employers, and establishment of the Social Security program of retirement benefits and unemployment and disability insurance. In his great work *The General Theory of Employment, Interest, and Money* (1936), the liberal English economist John Maynard Keynes introduced an influential economic theory that argued that government management of the economy could smooth out the highs and lows of the business cycle to produce more or less consistent growth with minimal unemployment.

Keynesian economics

Postwar liberalism to the 1960s. Liberalism, in strategic alliance with Soviet Communism, ultimately triumphed over Fascism in World War II, and liberal democracy was successfully reestablished in West Germany, Italy, and Japan. As western Europe, North America, and Japan entered a period of steady economic growth and unprecedented prosperity after the war, attention shifted to the institutional factors that prevented such economies from fully realizing their productive potential, especially during periods of mass unemployment and depression. Great Britain, the United States, and other Western industrialized nations committed their national governments to promoting full employment, the maximum use of their industrial capacity, and the maximum purchasing power of their citizenry. The old rhetoric about "sharing the wealth" gave way to a concentration on growth rates, as liberals—inspired by Keynes—used the government's power to borrow, tax, and spend not merely to counter contractions of the business cycle but to encourage expansion of the economy. Here, clearly, was a program less disruptive of class harmony and the basic consensus essential to a democracy than the old Robin Hood method of "taking from the rich and giving to the poor."

A further and final expansion of social-welfare programs occurred in the liberal democracies during the postwar decades. Notable measures were undertaken in Britain by the Labour government of Prime Minister Clement Attlee (1945–51) and in the United States by the Democratic administration of President Lyndon B. Johnson (1963–68) as part of his "Great Society" program of national reforms. These measures created the modern "welfare state," which provided not only the usual forms of social insurance but also pensions, unemployment benefits, subsidized medical care, family allowances, and government-funded higher education. By the 1960s social welfare was thus provided "from the cradle to the grave" throughout much of western Europe and in Japan, Canada, and the United States.

The liberal democratic model was adopted in Asia and Africa by most of the new nations that emerged from the dissolution of the British and French colonial empires in the 1950s and early '60s. The new nations almost invariably adopted constitutions and established parliamentary governments, believing that these institutions would lead to the same freedom and prosperity that had been achieved in Europe. The results, however, were mixed, with genuine parliamentary democracy taking root in some countries but succumbing in many others to military or socialist dictatorships.

CONTEMPORARY LIBERALISM

The revival of classical liberalism. The three decades of unprecedented general prosperity that the Western world experienced after World War II marked the high tide of modern liberalism. But modern liberalism was unprepared to cope with the slowing of economic growth that gripped most Western nations beginning in the mid-1970s. By the end of that decade economic stagnation, combined with the cost of maintaining the social benefits of the welfare state, pushed governments increasingly toward politically untenable levels of taxation as well as mounting debt. Equally troubling was the fact that the Keynesian economics practiced by many governments began to lose its effectiveness. The use of government spending to stimulate economic growth was causing increased inflation and producing ever-smaller declines in unemployment rates.

With modern liberalism seemingly powerless to boost stagnating living standards in mature industrial economies, the more energetic response to the problem turned out to be a revival of classical liberalism. The intellectual foundations of this revival were primarily the work of the Austrian-born British economist Friedrich von Hayek and the American economist Milton Friedman. One of Hayek's greatest achievements was to demonstrate, on purely logical grounds, that a centrally planned economy is impossible. He also famously argued, in his work *The Road to Serfdom* (1944), that interventionist measures aimed at the redistribution of wealth lead inevitably to totalitarianism. Friedman, as one of the founders of the modern monetarist school of economics, held that the business cycle is determined mainly by money supply and interest rates, rather than by government fiscal policy—contrary to the long-pre-

Monetarism

vailing view of Keynes and his followers. These arguments were enthusiastically embraced by the major conservative political parties in Britain and the United States, which had never abandoned the classical liberal conviction that the market, for all its faults, guides economic policy better than governments do. Revitalized conservatives achieved power with the lengthy administrations of British Prime Minister Margaret Thatcher (1979–90) and U.S. President Ronald Reagan (1981–89). Their ideology and policies, which properly belong to the history of conservatism rather than liberalism, became increasingly influential, as illustrated by the official abandonment of socialism by the British Labour Party in 1995 and by the cautiously pragmatic policies of U.S. President William J. Clinton in the 1990s.

Civil rights and social issues. Modern liberalism remains deeply concerned with reducing economic inequalities and helping the poor, but in recent decades it also has tried to extend individual rights in new directions. The concept of rights always had been used by liberals to argue against tyranny and oppression, but in the later 20th century it was massively expanded and became the most common way of formulating political demands. The prototypical mass movement in this regard was the American Civil Rights Movement of the 1950s and '60s, which resulted in legislation forbidding most forms of discrimination against a large African-American minority and fundamentally altered the climate of race relations in the United States. In the 1970s similar movements arose demanding equal rights for women, gays and lesbians, the physically handicapped, and other minorities or disadvantaged social groups. Thus liberalism historically has sought to foster a plurality of different ways of life, or different conceptions of the "good life," in society by protecting the rights and interests of first the middle class and religious minorities, then the working class and the poor, and finally blacks, women, homosexuals, and the physically or mentally disabled.

Liberalism has influenced the changing character of Western society in other ways as well, though its contribution in this regard has not always been distinguishable from the effects of modernization, technological change, and rising standards of living. For example, the abolition in most developed countries of traditional restrictions on contraception, divorce, abortion, and homosexuality was inspired in part by the traditional liberal insistence on individual choice. In similar fashion, the liberal emphasis on the right to freedom of speech has led to the loosening of inherited restrictions on sexual content and expression in works of art and culture. Indeed, liberalism proved so successful in enlarging a variety of personal freedoms that, to some critics, liberal values appeared to be eroding the strictures of morality and dissolving the traditional bonds of family and religion.

CONCLUSION

Liberalism survived the powerful totalitarian challenge of Fascism in the 1930s and '40s, and, with the collapse of the Soviet Union and the fall of its satellite regimes in eastern Europe in 1989–91, liberalism triumphed over its last remaining major ideological enemy, Soviet-style Communism. But today's liberals, sobered by the tragic events of the 20th century and chastened by abundant evidence of the defects in and limitations of human nature, no longer share their 19th-century predecessors' naive confidence in human rationality, human perfectibility, and the inevitability of progress. They are now more likely to agree with those who warn that human nature is ineradicably flawed than with those who hope to apply scientific methods to the solution of society's problems. Nevertheless, the continuing commitment of liberals to social reform suggests a persistent optimism and a belief that human beings can control their fate and build a better world. (H.K.G./K.Mi./Ed.)

Conservatism

Continuity and stability

The term conservatism denotes a preference for institutions and practices that have evolved historically and that are thus manifestations of continuity and stability. Political thought, from its beginnings, contains many strains that can be retrospectively labelled conservative, but it was not until the late 18th century that conservatism began to develop as a political attitude and movement reacting against the French Revolution of 1789. The noun seems to have been first used after 1815 by French Bourbon restorationists such as François-René, vicomte de Chateaubriand. It was used to describe the British Tory Party in 1830 by John Wilson Croker, the editor of *The Quarterly Review;* and John Calhoun, a formulator of conservative minority rights against majority dictatorship in the United States, also used the term in the 1830s. The generally acknowledged originator of modern, articulated conservatism (although he never employed the term) was the British parliamentarian and political writer Edmund Burke in his essay *Reflections on the Revolution in France* (1790). Pro-parliamentarian opponents of the French Revolution, such as Burke, believed that the violent, untraditional, and uprooting methods of the Revolution outweighed and corrupted its liberating ideals. More authoritarian opponents, such as the polemicist and diplomat Joseph de Maistre, also rejected the ideals themselves. The general revulsion against the course of events in France provided conservatives with an opportunity for restoring the pre-Revolutionary traditions, and a sudden flowering of more than one brand of conservative philosophy followed.

CONSERVATIVE ATTITUDES

Because Burke's case against radicalism and revolution has also influenced liberals, there is often no sharp distinction between liberals and conservatives in action. In philosophy, however, conservatism has maintained certain sharply nonliberal assumptions about human nature.

Whether intentionally or unconsciously, whether literally or metaphorically, for example, conservatives tend to assume in politics the Christian doctrine of man's innate original sin, and herein lies a key distinction between conservatives and liberals. Men are not born naturally free or good (conservatives assume) but are naturally prone to anarchy, evil, and mutual destruction. What the 18th-century French philosopher Jean-Jacques Rousseau denounced as the "chains" that hinder man's "natural goodness," are for Burkeans the props that make man good. These "chains" (society's traditional restrictions on the ego) fit man into a rooted, durable framework, without which ethical behaviour and responsible use of liberty are impossible.

Conservatism as a state of mind

The conservative temperament may be, but need not be, identical with conservative politics or right-wing economics; it may sometimes accompany left-wing politics or economics. Regardless of a conservative's politics or economics, however, it can be said that two characteristics of the conservative temperament are: a distrust of human nature, of rootlessness, of untested innovations; and a corresponding trust in unbroken historical continuity and in traditional frameworks within which human affairs may be conducted. Such a framework may be religious or cultural or may be given no abstract or institutional expression at all. In relation to the latter aspect, many authorities on conservatism—a minority in France and a majority in England—consider conservatism an inarticulate state of mind and not at all an ideology. Liberalism argues; conservatism simply *is.* When conservatism becomes ideologized, logical, and self-conscious, then it resembles the liberal rationalism that it opposes. According to this British approach, logical deductive reasoning is too doctrinaire, too 18th century. Whereas the liberal and rationalist mind consciously articulates abstract blueprints, the conservative mind unconsciously incarnates concrete traditions. And, because conservatism embodies rather than argues, its best insights are almost never developed into sustained theoretical works equal to those of liberalism and radicalism.

Conservatism is often associated with some traditional and established form of religion. After 1789, the appeal of religion redoubled for those craving security in an age of chaos. The Roman Catholic Church, because its roots are in the monarchic Middle Ages, has appealed to more conservatives than any other religion. Himself a Church of England Protestant, Burke praised Catholicism as "the most effectual barrier" against radicalism. But conservatism has had no dearth of Protestant and strongly anticlerical adherents also.

Conservatives typically view society as a single organism

and condemn as "rationalist blueprints" the attempts of progressives to plan society in advance from pure reason instead of letting it evolve naturally and unconsciously, flowering from the deep roots of tradition. They dismiss a liberal society as "atomistic," meaning composed of disrupted elements held together merely mechanically. A society, they argue, has to be rendered whole by religion, idealism, shared historical experiences, commitment to its long-standing political institutions, and by the emotions of reverence, cooperation, and loyalty; a society, they believe, can, to the contrary, be rendered atomistic by materialism, class war, excessive laissez-faire economics, greedy profiteering, overanalytical intellectuality, subversion of shared institutions, insistence on rights above duties, and by the emotions of skepticism and cynicism. Except for the German Romantic school, conservatives do not carry their conceptions of the organic wholeness of society to the extreme at which the individual becomes nothing, society everything, for they recognize that, at that extreme, one no longer has conservatism but totalitarian statism.

VARIETIES OF CONSERVATISM

The Burkean foundations. Burke did more than any other thinker to turn the intellectual tide from a rationalist contempt for the past to a traditionalist reverence for it. An Irishman, he loved England, including its established Anglican Church and its nobility, with an outsider's passion. In 1765 he became private secretary to Charles Watson-Wentworth, 2nd marquess of Rockingham, the head of the less liberal wing of the Whig Party. Against the untraditional tyranny of George III, Burke defended the American Revolution of 1776, which he viewed as being in defense of traditional liberties, but attacked the radical French Revolution of 1789 as tyranny by mobs and deracinated theorizers. At a time (1790) when the French Revolution still seemed a bloodless utopia, he predicted its later phase of terror and dictatorship, not by any lucky blind guess but by an analysis of its devaluation of tradition and inherited values.

Burke's conception of the social contract

Indeed, the core of Burke's thought and of conservatism is fear of rootlessness. Rousseau's *Social Contract* of 1762 had favoured a contract merely among the living, to arrange government for their mutual benefit. Burke, instead, argued:

> Society is indeed a contract . . . [but] as the ends of such a partnership cannot be obtained in many generations, it becomes a partnership not only between those who are living, but between those who are living, those who are dead, and those who are to be born. . . . Changing the state as often as there are floating fancies, . . . no one generation could link with the other. Men would be little better than the flies of a summer.

Burke's veneration of the past may be contrasted with the rationalist hostility of Karl Marx, the most influential social critic of modern times: "The legacy of the dead generations weighs like a nightmare upon the brains of the living." But for Burke the contract is with "the future" as well as with the past, and he thus urges improvement, as long as it is evolutionary: "A disposition to preserve and an ability to improve, taken together, would be my standard of a statesman."

Burke was defending not conservatism in the abstract but, rather, one concrete instance of it, the unwritten British constitution. His arguments, however, were not always consistent. Sometimes he justified that constitution by "natural rights"; more often by "prescriptive right." Natural rights meant a universal code external to any given constitution; prescriptive right, a local code authoritative (prescriptive) by virtue of its age and its links with the past, which are prima facie evidence of its value. Sometimes he argued that natural rights preceded the constitution and gave it "latent wisdom." But, when arguing against French rationalists, who would justify their own revolutionary constitution by natural rights, he argued instead, and more typically:

> Our constitution is a prescriptive constitution . . . [whose] sole authority is that it has existed time out of mind . . . without any reference whatever to any other more general or prior right.

Burke shocked his century by his brutal frankness in defending "illusions" and "prejudices" as socially necessary. In doing so, however, he was, in fact, being not so much a cynic as one of the few old-fashioned Christians among 18th-century intellectuals. He was an old-fashioned Christian in the sense of believing man innately depraved, innately steeped in original sin, and incapable of bettering himself by his feeble reason. So defined, man could be tamed only by following an ethically trained elite and by education in "prejudices," such as family, religion, and aristocracy. He called landed aristocrats "the great oaks" and "proper chieftains," provided they tempered their rule by a spirit of timely reform from above and remained within the constitutional framework. He defended the Church of England for its political as well as its religious function, "To keep moral, civil, and political bonds, together binding human understanding."

Coleridge and Wordsworth. After Burke, the English poets Samuel Taylor Coleridge and William Wordsworth were significant figures in the formulation and expression of conservative sentiment. They began, however, as utopian liberals supporting the French Revolution. Wordsworth spoke for a whole generation of European intellectuals with his famous salute to the new dawn in France: "Bliss was it in that dawn to be alive, but to be young was very heaven." Disillusionment followed, and Coleridge and Wordsworth reacted against liberalism and rationalism and turned to traditional monarchy and the Church of England.

In 1798 Wordsworth and Coleridge jointly published their book of poems, *Lyrical Ballads,* marking the revolt of the human heart against abstract 18th-century rationalists and thereby helping to create a new philosophical climate. Conservatism was permanently influenced by Coleridge's prose works: *Lay Sermons,* 1816–17; *Biographia Literaria,* 1817; *Philosophical Lectures,* 1818–19; *Aids to Reflection in the Formation of a Manly Character, on the Several Grounds of Prudence, Morality, and Religion,* 1825; and his various *Letters* and *Specimens of Table Talk.* His public lectures exercised an indirect influence by molding the minds of university students who later became national leaders.

Coleridge's views on social classes

According to Coleridge, society divided its functions among different "class orders." Each class had its valuable function, but this did not necessarily include the right to vote and rule. That right was best left to an ethically trained aristocracy, functioning within the strict lawful limits of Parliament. All classes, Coleridge argued, must cooperate harmoniously within the organic unity of the constitution. His greatest influence on practical politics was through his disciple Benjamin Disraeli, later to be Conservative prime minister, and his disciple's disciple, Sir Winston Churchill. Coleridge considered businessmen often subversive, not conservative; they allegedly gnawed at the foundations of Christian monarchy by substituting a newfangled, un-Christian religion known as economic profit. Thus Coleridge, defining "shopkeepers" as "the least patriotic and the least conservative" class, fought against the Whig Reform Bill of 1832, which made "hucksters" the dominant voting group.

Maistre and Latin conservatism. It would convey an unbalanced picture of conservatism to present only the moderate and British brand founded by Burke and to omit the more extreme and Latin brand founded by Maistre (died 1821). Whereas Burkean conservatism is evolutionary, the conservatism of Maistre is counterrevolutionary. Both favour tradition against the innovations of 1789, but their traditions differ: the former fights against 1789 for the sake of traditional liberties, the latter for the sake of traditional authority. The former is not authoritarian but constitutionalist—and often parliamentary—whereas the latter, in its stress on the authority of some traditional elite, is often justifiably called not conservative but reactionary. To call it totalitarian, however, would be to go much too far, for its authority does not try to be "total," in the sense of taking over the total personality, the total culture, but is restricted to politics—and sometimes also religion. The distinction between the authoritarian and the totalitarian

separates even the most reactionary conservative from the totalitarian Nazis and Communists.

After the breakdown of the French Revolution, Maistre became the most influential philosophical spokesman for the *ancien régime*. Against the slogan "liberty, equality, fraternity," he seemed almost personally to embody the slogan "throne and altar." His program consisted of a restoration of hereditary monarchy, but a more religious and less frivolous monarchy than before. He was an international refugee after the French, during the Revolution, invaded his native Savoy—then a French-speaking province of the Italian-speaking monarchy of Piedmont-Sardinia. He became for 14 years Sardinian ambassador to Russia, where his restorationist faith was strengthened by the example of the absolute monarchy still functioning there.

Maistre on the role of monarchy

Both restorationist and evolutionary conservatives defended monarchy as a social cement needed to hold society together, to keep it "organic," not "atomistic." But, while the Maistre school (key source of conservative thought in Spain and Italy as well as France) defends monarchy as absolute, the evolutionary British school defends it merely as being "pragmatic"; that is, useful. Maistre and many continental monarchists carried their belief in the monarchy to the extreme of demanding "love" even for an "unjust" ruler, earthly or heavenly:

> We find ourselves in a realm whose sovereign has proclaimed his laws. . . . Some . . . appear hard and even unjust . . . What should be done? Leave the realm, perhaps? Impossible: the realm is everywhere. . . . Since we start with the supposition that the master exists and that we must serve him absolutely, is it not better to serve him, whatever his nature, with love than without it?

This chain of authoritarian reasoning reached its climax in a logical if inhuman paradox: "The more terrible God appears to us . . . the more our prayers must become ardent. . . ." Cruel as these arguments sound, the motive of the personally mild Maistre was humane: revolts against cruel authority would inflict even crueler sufferings on mankind. He drew from the French Revolution the lesson that submission to traditional authority, though admittedly a bitter pill, was Europe's cure for a still more bitter chaos.

Maistre's politics were a theological drama in which "order" (his key concept) was angelic, "chaos" diabolic, and "revolution" original sin. Seduced by the glittering *Social Contract* of Rousseau, giddy and inexperienced nations might lust after democracy or a plebeian Bonapartist dictatorship. But they would come to a perfectly dreadful end, which would serve them right for provoking the wages of sin: "Because she [Europe] is guilty, she suffers" (1810). From suffering, Maistre argued, Europe would learn that the purest order is a fatherly Christian monarchy. Even kings must avoid rocking the boat of order with liberal "innovations": Europe must "suspect" the word "reform." In *Du Pape* (1817; "Concerning the Pope"), he analyzed "order" further: its hierarchical pyramid logically required one supreme apex. That apex must be no earthly monarch, of which there were so many, but the union of earthly and spiritual power in the papacy.

The vast extent of the instability following the French Revolution surprised even its supporters, and the problem of how to restabilize society emerged as one of some practical importance. According to Maistre's *Soirées de Saint-Pétersbourg* (left unfinished 1821; "Evening Conversations in St. Petersburg"), the solution was more faith and more police. That combination he summed up in his own frank formula: "the pope and the executioner." The pope was the positive bulwark of order: he gave faith. The executioner was the negative bulwark: he suppressed disorder. Himself an intellectual, Maistre indicted intellectuals as "rebellious" and "insolent" fomenters of disorder.

Maistre, this very secular exalter of clericalism, resembled not the Church Fathers but the very rationalists he attacked. He arrived at his glorification of unreason and of divine authority not by mystic intuition—not even by unthinking acceptance of traditional authority—but by using his own mind independently, rationally, and with steps of deductive logic. Though Maistre would never have admitted it, he might be characterized as the last abstract rationalist of the whole Voltairean Age of Reason. Even more than the rationalist Voltaire and as much as the rationalist Jacobins, Maistre believed in pure and absolute ideas, although his idea was absolute authority rather than absolute reason. In Maistre the destructive deductive logic of the 18th century was carried so far that it destroyed even itself—pure reason committing suicide for the sake of pure order.

This division into Burke and Maistre wings does not mean both were equal in importance or influence. No work of Maistre or any other anti-Jacobin has approached the influence of Burke's classic essay. Burke, above all, was the first to formulate the rebuttal to the French Revolution; his arguments were borrowed, sometimes word for word, by all later conservatives, including the restorationists. Maistre's rigid hierarchical conservatism is in the latter part of the 20th century dying out, whereas Burke's more flexible brand is stronger than ever, permeating all parties of the West, emphatically including democratic Socialists with their increasing stress, in Great Britain and Germany, on what a Fabian Socialist has called, in good Burkean language, "the inevitability of gradualness."

Later French conservatism

French conservatism after Maistre presents a diversified range of views, from the thought of Charles Maurras, the far-right editor of *L'Action Française* who seemed more fascist than conservative and became a Nazi collaborator, to the anti-authoritarian Alexis de Tocqueville, author of *Democracy in America* (1835–40) and the most Burkean French critic of the Revolution and of plebiscitarian mass democracy. To some extent, however, Tocqueville, an evolutionary parliamentarian, can also be regarded as a liberal thinker. In between Maurras and Tocqueville come the great anti-Jacobin Hippolyte-Adolphe Taine; the philosophical novelist Maurice Barrès, more a nationalist than anything else but conservative in his stress on organic roots; and Louis-François Veuillot, the editor after 1843 of the newspaper *L'Univers Réligieux* and a clerical restorationist who ably readapted Maistre to the industrial modern world. An influential right-wing extremist, less clerical and more statist than Maistre and Veuillot, was Louis-Jacques-Maurice de Bonald, the apologist for Napoleon's empire and then for the Bourbon Restoration.

Metternich and the Concert of Europe. The problems posed by the widespread social unrest of the Revolutionary and Napoleonic periods and their aftermath, and the insecurity of governments in the face of demands for constitutions and liberal reforms, provoked a reaction of more immediate and far-reaching consequence than the writings of conservative theorists. During the period 1815–48, Prince Metternich, a major influence in Austria and in Europe generally, devoted his energies to erecting an antirevolutionary chain of international alliances throughout Europe in order to protect the multinational empire that he administered.

Metternich viewed the liberal revolutions of the 1820s and 1830s in Italy, Spain, and Germany as being unhistorical and unrealistic. Liberals were trying to transplant from England free institutions, which had no historic roots on the Continent. He retorted with Burkean arguments about the need for old roots and orderly organic development. Hence, his sarcastic comments on the liberal revolutions in Naples and elsewhere:

> A people who can neither read nor write, whose last word is the dagger—fine material for constitutional principles! . . . The English constitution is the work of centuries. . . . There is no universal recipe for constitutions.

Though his repressive Carlsbad Decrees of 1819 infringed inexcusably on basic liberties, his attitude was not always so negative. Just before his fall in 1848, he was at last winning acceptance from the archdukes of his sincere, thoughtful, and practical plan (postponed too long by the reactionary emperor Francis I) to convoke delegates from all the provincial estates to a representative body in Vienna.

Metternich was a dominating figure at the Congress of Vienna, the international peace conference of 1815 after the Napoleonic Wars. The Vienna peace was based on certain conservative principles shared by the Austrian delegate

Conservative objectives following the Napoleonic Wars

Metternich, the British delegate Robert Castlereagh, the French delegate Talleyrand, and the formerly liberal Russian tsar Alexander I. These principles were conservatism, in reaction against Revolutionary France; traditionalism, in reaction against 25 years of rapid change; legitimism (the principle of hereditary monarchy as the only lawful rule); and restoration (the principle of restoring the kings ousted after 1789).

The European great powers also aimed at the enforcement of peace by subsequent conferences between kings, and those subsequent conferences gave rise to a period of international cooperation known as the Concert of Europe. As liberal democrats correctly pointed out, the weakness of that first successful attempt at a "United Nations" was its narrowly aristocratic base. But it did achieve the positive function—and important precedent—of peacefully arbitrating several disputes. The debit of the conservative Concert of Europe was its bigoted suppression of democratic social progress.

Goethe's spiritual conservatism. Johann Wolfgang von Goethe was Germany's greatest dramatist, poet, and personality. In his youthful "storm and stress" period of the 1770s, Goethe went through a phase of revolt and of nationalism. In his old age, however, he became Germany's greatest cultural influence for classical balance and for antinationalist cosmopolitanism, influencing many outside Germany, including, in England, Coleridge. In 1815 Goethe and Metternich both took pride in being "good Europeans," not German nationalists. After a friendly personal conversation with Metternich, Goethe wrote that Metternich "inspires with the assurance that reason, reconciliation, and human understanding will lead us out of present chaos." Later, in 1830, Goethe urged a mature synthesis between a conservative framework and liberal goals:

> The genuine liberal tries to achieve as much good as he can with the available means to which he is limited; but he would not use fire and sword to annihilate the often inevitable wrongs. Making progress at a judicious pace, he strives to remove society's deficiencies gradually without at the same time destroying an equal amount of good by violent measures. In this ever-imperfect world he contents himself with what is good until time and circumstances favor his attaining something better.

His rhymed credo "Nature and Art" (1802) expressed his conservative and classic stress on voluntary submission to law: "Only in self-restriction does the master reveal himself. And only law can give us liberty." His political drama *Die natürliche Tochter* (1803; *The Natural Daughter*) reflected his hostility to the French Revolution, radicalism, and mass movements. Much quoted by classicists, such as the United States' Irving Babbitt, was Goethe's definition: "The classical I call the healthy and the romantic the diseased." Yet his *Faust* drama (*Part I* published 1808, *Part II* 1832) retained the liberal-minded stress of his younger days on constant change, "constant striving," as salvation. His most unique achievement consisted of his being, so to speak, self-invented. By sheer strength of character, he remolded his naturally revolutionary and romantic temperament into what the world accepted as a conservative and classicist temperament.

Savigny and Ranke

Perhaps Germany's most mature conservative thought came from her great historians. Friedrich Karl von Savigny (died 1861) and Leopold von Ranke (died 1886) were outstanding as pupils of Burke in their reverence for history as organic growth. Savigny stressed that custom, operating over centuries, creates its own framework. On custom, Savigny founded an entire science of historical jurisprudence, denying the abstract, liberal "rights of man." Similarly, Ranke saw every society in terms of its own unique evolution. He opposed the universal generalizations of the 18th-century Enlightenment; every people, he wrote, "is related directly to God" in its own concrete way.

Tsarist and Dostoyevskyan conservatism. Whereas Western conservatism arose from reactions to the French Revolution, Russian tsarist conservatism had different and older origins. The practice of the absolute Tatar khans and the theory of Byzantine caesarism combined to produce an un-Western elephantiasis of autocracy. Nevertheless, two antiliberal traditionalists of Russia made such an impact on the West—the first by politics, the second by art—that their mention is indispensable: Konstantin Pobedonostsev and Fyodor Dostoyevsky. The former was the tutor and chief ideologist of two tsars (Alexander III and, until the Revolution of 1905, Nicholas II). His book *Reflections of a Russian Statesman* (1898) denounced free press, trial by jury, parliamentary government, secular education, skepticism toward the divine mission of tsars, and, above all, intellectuals.

Dostoyevsky's disillusionment with his youthful radicalism resembled Coleridge's in its psychological as well as literary consequences. Both turned to an organic, religious, and monarchic society, to which they paid more homage via literature than via politics. Dostoyevsky attacked Socialism, liberalism, materialism, and atheism. He preached Greek Orthodox tsarism, Slavic traditionalism, and the redemption of mankind by "Holy Russia." His novel *The Possessed* (1871–72) pictured the idealistic ends of Socialists as corrupted by their terroristic means, and he boasted somewhat fawningly to Alexander III about the book's effectiveness against radicals. His novel *The Brothers Karamazov* (1880) contrasted a dry Western rationalism with a more deeply moving Russian mysticism. To the end he retained from his young Socialist days his characteristic compassion for what he called "the insulted and injured"; only now he expressed this in the more spiritual creed of Christian love. What influences many modern readers so compellingly is not his political but his cultural conservatism, exalting vision beyond external material progress.

American conservatism. The American Revolution owed many of its ideals to Burke's interpretation of the British heritage of 1688, the heritage of mature self-government. Burke favoured the Revolution as defending the traditional rights of freeborn Englishmen against newfangled royal usurpations. In that sense, one might describe it not as the Revolution but as the "Conservation" of 1776.

In *The Rights of the British Colonies Asserted and Proved* (1764) the American spokesman James Otis typically argued that the demand for no taxation without representation was an old British tradition. America, he said, was conserving "the British Constitution, the most free one on earth." "We claim nothing," added George Mason of Virginia, "but the liberty and privileges of Englishmen." Almost all other revolutions, colonial or otherwise, have been radical in the sense of demanding new or increased liberties and a new order. In contrast, the American demand of July 6, 1775 (*Declaration of the Causes & Necessity of Taking Up Arms*), was for conserving old liberties and the old order: "in defence of the freedom that is our birth right and which we ever enjoyed until the late violation of it." Such words promulgated no democracy, no abstract "Rights of Man"; rather, they promulgated what Burke called "prescriptive right. . . . considering our liberties in the light of an inheritance." Despite important exceptions, which should not be minimized, it was not until the election of the more truly "revolutionary" Andrew Jackson (1828) that the democratic doctrines of the pamphleteer Thomas Paine gained solid roots in the United States, dividing the nation between conservative and progressive traditions. Paine was the man whom the Burkean John Adams (president 1797–1801) came to loathe most—for eternally sloganizing about apriorist utopias. A leading historian, Daniel Boorstin, has observed in *The Genius of American Politics* (1953):

> The ablest defender of the Revolution—in fact, the greatest political theorist of the American Revolution—was also the great theorist of British conservatism, Edmund Burke. . . . Ours was one of the few conservative colonial rebellions of modern times.

Conservative doctrines of Hamilton and Madison

The spirit of the United States was partly molded by two masterpieces of Burkean conservatism, both published in 1787–88: *The Federalist*, by Alexander Hamilton, James Madison, and John Jay, and *Defence of the Constitutions of Government of the United States of America*, by John Adams. The achievements attributed by historians to the *Federalist* papers exceed those of any other series of newspaper articles in history, for they helped forge national

unity during a separatist crisis. In the context of Shays's Rebellion of 1786 against the judiciary, they saved government by law from government by mob and established minority rights against majority dictatorship. They based American liberty on the Burkean principle of historical roots, prescriptive right, and judicial precedent instead of on vague grand rhetoric about democratic utopias and the masses. Similar in thought and richer in historical background was the *Defence* by Adams, one of the most penetrating analyses of self-government ever written.

The U.S. Constitution was drawn up in Philadelphia by the U.S. Constitutional Convention of 1787. The objectives of many liberal democrats were: easy amendment; facilities for mass pressure and rapid change; unchecked popular sovereignty; universal manhood suffrage; a single parliamentary body; and the basing of liberty on a long list of universal a priori abstractions, such as Burke later criticized in the French Declaration of the Rights of Man and of the Citizen. But in the Constitution of 1787 the Federalists foiled each of these objectives. They made amendments slow and difficult, greatly reduced the number of voters by property restrictions, created a congress of two parliamentary bodies, and based liberty primarily, though not entirely, on the concrete, inherited precedents of British tradition. Except for the House of Representatives (a sop to democrats), the main cogs of government—president, Senate, justices—were not to be chosen directly by the people but, respectively, by the electoral college, state legislatures, and appointment, and not until 1913 did an amendment eliminate this intentionally undemocratic election of senators. The judicial branch (Supreme Court) continues to be a nonelective, nonremovable elite not responsible to democratic majorities. Yet it can veto as unconstitutional measures passed by a democratic majority of the two elective, removable branches of Congress.

The American Founding Fathers adopted a conservative constitution in reaction against current mob excesses and against the democratic–utopian rhetoric of the earlier Declaration of Independence (drawn up by Thomas Jefferson) with its grand abstractions about "life, liberty, and the pursuit of happiness." Yet the Constitution was the Burkean, not the reactionary brand of conservatism. Thus it defeated not only the liberal objectives but also the more extreme conservative ones, including a hereditary, titled aristocracy and Hamilton's notion of a president for life with absolute veto power.

The United States' only consistently conservative party was the Federalist Party of John Adams and Alexander Hamilton. Hamilton was perhaps too much the reckless commercial adventurer to be classified under conservative or any other principles, but Adams remains the closest New World equivalent to Burke. After the death of the Federalist Party in the early 1800s, two mutually hostile kinds of political conservatism emerged: that of the urban New England Brahmins and that of the Southern semi-feudal landowners. The latter received their most persuasive defense in the famous *A Disquisition on Government and Discourse on the Constitution and Government of the United States* of Calhoun, the closest New World equivalent to Maistre. This more extreme, very regional Calhoun conservatism is still influential in much of the American south, typically cutting across Democrat or Republican party lines, and is still alien to New England conservatism.

John Calhoun

Modern U.S. political parties, being pragmatic alliances of geographic patronage groups rather than matters of doctrine, cannot realistically be classified under "isms." It is nearer to reality to look for conservatism, instead, in the indirect diffusion—cutting across all party lines—of the above described restraining principles of the Constitution.

MODERN CONSERVATISM

The 19th and, particularly, the 20th century (that is, the period since the 18th-century Enlightenment) have in many ways been antithetical to conservatism, both as a political philosophy and as a program of particular parties identified with conservative interests. As described above, the consciously articulated conservatism of Burke was formulated in reaction to the French Revolution; similarly, the anti-liberal, anti-revolutionary policy that was a major factor in European international relations during the Metternich period (1809–48) was a reaction to the political discontent aroused by demands for liberal reforms and constitutions. The Enlightenment, in fact, had resulted in the propagation of certain attitudes and ideas that were to have far-reaching political consequences during the succeeding centuries, the most significant of which were a belief in the possibility of improvement in the human condition and a concomitant disposition to tamper with or discard existing institutions or practices in pursuit of that goal—a disposition that conservatives, as noted above, have regarded as "rationalist." Such rationalist politics embrace a broad segment of the political spectrum, including much of liberal reformism, socialism of the welfare-state or mixed-economy variety characteristic of western Europe, and Marxist socialism. The changes that have been wrought under the banner of rationalist politics have thus been immense and point to what has been described as a dilemma of modern conservatism—the extent to which, in face of constant rationalist innovation, conservatives may be forced to adopt a merely defensive role, so that the political initiative lies always in the other camp.

The responses of conservatives to this predicament have naturally varied considerably in differing political contexts; an account of some of these responses is given below. An analysis of the role of conservatism in contemporary politics, however, cannot be confined merely to an account of the programs of political parties identified with the conservative cause, for conservatism makes its influence felt in a variety of ways less direct than through expression in party platforms. Modern conservatism is in fact a pervasive force in the political life of those parliamentary democracies in which rationalist politics seem to hold sway—as well, of course, as in less liberal political climates.

Non-political manifestations of conservatism. Conservative influences operate indirectly (*i.e.*, other than via the programs of political parties) largely by virtue of the fact that, while man is undeniably a persistent innovator, there is also much in the human temperament that is naturally or instinctively conservative: among such conservative traits are the tendency to fear and avoid sudden change and the tendency to act according to habit. While these are traits of the individual, they may find collective expression in, for example, resistance to imposed political change and in a whole cluster of values that contribute to the formation and stability of a particular culture. The tendency for values to find expression in cultural forms and political institutions (the so-called pragmatism of the British, for example, in their unwritten constitution) constitutes a profound conservative influence in political life over and above any explicit articulation of particular conservative interests that may be undertaken by a political party, for it gives rise to practices and institutions that are products of a long process of social and political evolution and are closely related to other cultural factors, such as religion and property relationships. The existence of such cultural restraints on political innovation constitutes in all societies a fundamental conservative bias, the implications of which have been aphoristically expressed by an English commentator, F.J.C. Hearnshaw: "It is commonly sufficient for practical purposes if conservatives, without saying anything, just sit and think, or even if they merely sit." Mere inertia, however, has rarely sufficed to protect conservative values in an age dominated by rationalist dogma and by social change related to continuous technological developments. The conservative reaction, however, is best analyzed in specific political contexts. Historians, it may be noted, cannot safely agree on there being more than four great political parties of the 20th century deserving of the name: the Conservative Party of England, the Christian Democrats of Italy and of Germany, and the Liberal Democrats of Japan.

Conservative traits of human nature

Great Britain. In England, Disraeli's successor, Lord Salisbury, was prime minister in 1885, from 1886 to 1892, and from 1895 to 1902; Arthur Balfour succeeding him from 1902 to 1905. This longest era of Conservative rule was characterized by imperialism, high tariffs, and the gradual erosion of the party's working-class vote, which Disraeli had so far-sightedly nurtured by extending the

franchise to the workers in 1867. The party had thereby broadened its original class basis (landed aristocracy and established church) to outflank from below and above the new commercial class and its Liberal Party. It may be said that conservatism in Great Britain since Disraeli's time has veered between a passive and largely resigned acceptance of changes introduced by its Liberal and, later, Labour opponents and a more positive conservatism, the aim of which has been to foster a social environment in which the individual is encouraged to advance his own interests without undue hindrance from, or reliance on, the state—a policy descended from the liberal individualism of the 19th century, associated particularly with the Liberal Party. This positive conservatism of liberal individualism tinged with a strong sense of social conscience was given its earliest formulation by Disraeli, who combined a desire to mitigate harsh conditions suffered by the working class under conditions of unrestrained capitalism with a belief in the value of existing institutions such as the monarchy, the church, and the class system. Disraeli's foreign policy, which emphasized the need for Britain to act constructively as a "moderating and mediatorial" power and to maintain its interest in its empire, also reflected the view that conservatism must be a force shaping events rather than merely reacting to them. These three elements—the improvement of material conditions by both encouragement of individual initiative and timely reform of abuses, emphasis on the value of traditional institutions, and belief in the need for an active foreign policy—have been recurring themes of British conservatism in the 20th century. Later conservative thinkers have elaborated on the value of divergency of personality and attitudes, the role of property as an expression of individuality, and the central role of the family in providing a stable environment in which the individual may develop. In its less positive periods (as, for example, during the interwar period), conservatism in Britain has been identified with the defense of class privileges and of the status quo, an unconstructive opposition to socialism, and, during the 1930s, a deal-making commercialist approach to the rising Nazi menace.

British conservatism after Disraeli

In 1945–51 the Labour government of Clement Attlee created a modern welfare state in Britain whose general features, including the nationalization of major corporations and entire industries and the vast extension of social services, were later imitated by other European nations and, to a limited extent, by the United States. During the subsequent two decades the Conservatives, when in power, generally refrained from reversing these reforms, emphasizing instead their claim to be able to administer the welfare state more efficiently than the Labour Party. This era of liberal-conservative accomodation came to a dramatic close under the government of Margaret Thatcher (1979–90), whose energetic brand of conservatism stressed individual initiative, fierce anti-communism, and free-market economics.

Continental Europe. Conservatism elsewhere in western Europe was generally represented by two or more parties, ranging from the liberal centre to the moderate and extreme right. Three types of party may be discerned: agrarian parties (particularly in Scandinavia), Christian Democratic parties, and conservative parties strongly linked to big businesses. These categories are very general and are not mutually exclusive.

The Christian Democratic tradition

The Christian Democratic parties have the longest history, their predecessors having emerged in the 19th century to support the church and the monarchy against liberal and radical elements. After World War I, supporters of business became the predominant element in these parties. Clerical interests remained strongly represented in the Democrazia Cristiana (Christian Democrat Party) of Italy, which dominated governments in that country for four decades from 1945. This party never possessed a coherent policy, however, because it was little more than a disparate alliance of moderate and conservative interest groups. The Christian Democrats anchored a long series of governing coalitions with smaller centrist parties and the Italian Socialist Party. These coalitions, while often politically ineffective and increasingly corrupt, served to exclude the large Italian Communist Party from power throughout the Cold War. When the Soviet Union collapsed in 1991 and communism was no longer perceived as a threat to Europe, the Christian Democrats lost much of their support and soon were reduced to a minor party. The eclipse of the Christian Democrats coincided with the growth of other conservative and nationalist groups formerly outside the mainstream of Italian politics, such as the Northern League, which called for the creation of a federated Italian republic, and the National Alliance (until 1994 the Italian Social Movement), which many regarded as neofascist.

In Germany, a country divided between Catholics and Protestants, the church played a far less significant role in the main conservative party, the Christlich-Demokratische Union (Christian Democratic Union). After 1950, following an internal debate over economic and social questions, the party adopted a program that included advocacy of a free-enterprise economy and a strong commitment to maintain and improve social insurance and other welfare provisions. Illustrating the conservative temper of Germany's political climate since the end of World War II, the opposition Sozialdemokratische Partei Deutschlands (Social Democratic Party of Germany) has progressively eliminated the socialist content of its program, going so far as championing the profit motive in a party congress at Bad Godesberg in 1959. In power continuously from 1982 to 1998, the Christian Democrats presided over the unification of East Germany with West Germany following the collapse of Soviet-supported communist regimes across eastern Europe in 1989–90.

Unlike in Italy and Germany, moderate conservative opinion in France was not represented by a Christian Democratic party. Instead, a large proportion of French conservatives supported parties such as the Rassemblement pour la République (Rally for the Republic), which espoused a highly nationalistic conservatism based on the legacy of Charles de Gaulle, president of France from 1958 to 1969, or anti-immigration groups like the Front National (National Front), led by political firebrand Jean-Marie Le Pen (who, some would argue, is not so much conservative as reactionary or neofascist). Gaullist conservatism emphasizes tradition and order and aims at a politically united Europe under French leadership. Gaullists espouse divergent views on domestic social issues, however. The large number of Gaullist and non-Gaullist conservative parties, their lack of stability, and their tendency to identify themselves with local issues make it difficult to categorize these groups in simple terms.

In general, conservatism in Europe has been a pervasive political influence, finding expression in parties of very different character. These parties typically represent traditional middle-class values and oppose unnecessary state involvement in economic affairs and any radical attempts at income redistribution. They are also characterized by an absence of ideology and often by the lack of any well-articulated political philosophy.

Japan. The relationship between conservatism as an underlying bias related to psychological factors and cultural values and conservatism as an articulated political credo is illustrated by the history of party politics in Japan since its opening to Western influence in the middle of the 19th century. The political and social changes that took place following the Meiji Restoration (1868) were of major proportions, involving the abolition of feudal institutions and the introduction of Western political ideas such as constitutional government. But despite institutional innovations and the dislocations resulting from rapid industrialization, traditional loyalties and attitudes proved to be more important factors in shaping political developments.

Majority conservatism in Japan

Except for the period of intervention by the militarists during the 1930s and '40s, Japan has been ruled by conservatives almost continuously since the beginning of party politics in the 1880s. The conservative parties (the two most important of which merged to form the Liberal-Democratic Party in 1955) have been dominated by personalities rather than by ideology and dogma; and personal loyalties to leaders of groups within the party (factions) rather than commitment to policy have determined the allegiance of conservative members of the Diet.

The Liberal-Democratic Party is intimately linked with big business interests, and its policies are guided primarily by

the objective of fostering a stable environment for the development of Japan's free-enterprise economy; to this end, the party functions as a broker of conflicting business interests. Policy toward other Asian countries, national defense, and internal security are other conservative preoccupations.

The United States. Politics in the United States never quite conformed to the doctrinal patterns exhibited in continental Europe or even Britain, mainly because there was never a monarchy, an aristocracy, or an established church for conservatives to defend or for liberals to attack. The Federalists of the late 18th and early 19th centuries were conservative in their emphasis on order and security but were classic liberals in many other respects. The nearest thing to an American aristocracy was the wealthy plantation-owning class in the South before the Civil War. Members of this class generally favoured the rights of states against the power of the federal government, and prominent defenders of states' rights, such as John Calhoun, have properly been seen as conservative thinkers. This particular brand of conservatism eventually took on an antimodernist tone, as evidenced by the rise of the Southern Agrarians, whose 1930 manifesto, *I'll Take My Stand,* continues to inspire like-minded traditionalists in the United States today.

The Southern Agarians

But if there was relatively little explicit conservatism in the United States until well into the 20th century, the political history of the country has also been remarkably resistant to revolutionary radicalism. The American working class has generally shared the hopeful individualism of the middle class. As a result, the common view has been that the United States is a country of one basic political tradition: liberalism. For a long time it seemed that conservatism could not take root in a country founded on the liberal doctrines of the founding fathers.

This perception began to change in the wake of the New Deal, the economic relief program undertaken by the Democratic president Franklin D. Roosevelt in 1933 to help raise the United States out of the Great Depression. This program greatly expanded the federal government's involvement in the economy through the regulation of private enterprise, the levying of higher taxes on corporations and the wealthy, and the expansion of social-welfare programs. The New Deal was stubbornly opposed by the Republican Party, whose main supporters were big business, the wealthy, and prosperous farmers. Conservatives also objected to Roosevelt's 1937 proposal to reorganize the Supreme Court. Widely regarded as a court-packing scheme, the measure would have allowed the president to appoint up to six new justices, thereby ensuring a liberal majority and removing the threat that the Court would invalidate key New Deal legislation such as the Wagner Act and the Social Security Act.

Generally, with Democratic liberals moving to the left in endorsing a larger role for government, the Republicans clung to a 19th-century version of liberalism that called for the government to avoid interfering in the free market. This policy produced little success for Republicans at the polls. In matters of foreign policy, however, the Old Right, as these staunch conservatives were known, was powerful enough to prevent the United States from entering World War II until the Japanese attack on the U.S. naval base at Pearl Harbor, Hawaii, in 1941. By the time the Republicans regained the presidency in 1953, they had accepted most of the New Deal reforms and were preoccupied with battling communists at home and abroad.

In the first decades after the war, the United States, like Britain, gradually expanded social services and increased government regulation of the economy. In the 1970s, however, the postwar economic growth that Western governments had relied on to finance social-welfare programs began to slacken, just as Japan and other East Asian nations were finally attaining Western levels of prosperity. Whatever the causes of the West's economic stagnation, it became clear that liberal policies of governmental activism were incapable of solving the problem.

The "neoconservatives"

At this point a new group of mainly American conservatives, the so-called "neoconservatives," arose to argue that the chief factors discouraging economic growth were high levels of taxation and the government's intrusive regulation of private enterprise. Inspired by the work of free-market economists such as Friedrich von Hayek and Milton Friedman, they generally accepted a minimal welfare state—something an older breed of conservative would never have done—and fought simply for reductions in social benefits and government spending, lower taxes, and less government regulation of business. They also shared little of earlier conservatives' isolationist tendencies, protectionist impulses, and jealous regard for national sovereignty. Many of them argued that the United States had a right to intervene in the affairs of other nations in order to combat the influence of Soviet communism and to advance its own national interests; some even claimed that the United States had a duty to remake the non-Western world on the model of American democratic capitalism. In economic matters, neoconservatives were often hard to distinguish from classical liberals of the 19th century, who had likewise urged a minimum of government intervention in the economic life of nations. Among American political leaders, the chief representative of neoconservatism was Republican President Ronald Reagan (1981–89).

CONSERVATISM AT THE TURN OF THE 21ST CENTURY

Division, not unity, marked conservatism around the world at the end of the 20th century—this despite the defeat of conservatism's chief nemesis over the last half century, Soviet communism. But perhaps this fissure is not surprising. Anticommunism was the glue that held the conservative movement together, and without this common enemy the many differences between conservatives became painfully clear. In Europe, for example, conservatives split over such issues as the desirability of a united Europe, the benefit of a single currency, and the region's proper role in policing troubled areas like the Balkans and the Middle East. Conservatism was even more divided in the United States. Abortion, immigration, national sovereignty, and "family values" were among the issues that rallied supporters but divided adherents into various camps, from neoconservatives and "paleoconservatives" (descendants of the Old Right who saw neoconservatives as socially liberal and imperialistic in foreign affairs) to cultural traditionalists among groups such as the Christian Coalition and the Moral Majority. The camps battled one another as well as perceived enemies in what were termed the "Culture Wars" in the 1990s, and, through it all, each faction was convinced that it alone was carrying the true mantle of conservatism into the next millennium. (Pe.V./K.Mi./Ed.)

Contemporary debates

BIBLIOGRAPHY

General. WILLIAM OUTHWAITE and TOM BOTTOMORE (eds.), *The Blackwell Dictionary of Twentieth-Century Social Thought* (1993), is an excellent and comprehensive resource.

Some fundamental introductions to various socio-economic doctrines which put them in perspective with each other are BARRINGTON MOORE, JR., *Social Origins of Dictatorship and Democracy* (1966, reissued 1993); ROBERT D. PUTNAM, *The Beliefs of Politicians: Ideology, Conflict, and Democracy in Britain and Italy* (1973); WOLFGANG LEONHARD, *Three Faces of Marxism: The Political Concepts of Soviet Ideology, Maoism, and Humanist Marxism* (1974; originally published in German, 1970); IMMANUEL WALLERSTEIN, *The Modern World-System,* 3 vol. (1974), a stimulating historical analysis of the origins of the world's political and economic systems and their relationship to each other; CHARLES E. LINDBLOM, *Politics and Markets: The World's Political Economic Systems* (1977); ERIK OLIN WRIGHT, *Class, Crisis, and the State* (1978); KENNETH MINOGUE, *Alien Powers: The Pure Theory of Ideology* (1985); JOSEPH A. SCHUMPETER, *Capitalism, Socialism, and Democracy,* 6th ed. (1987); NEIL J. SMELSER (ed.), *Handbook of Sociology* (1988); GREGORY M. LUEBBERT, *Liberalism, Fascism, or Social Democracy* (1991); ROBERT EATWELL and ANTHONY WRIGHT (eds.), *Contemporary Political Ideologies* (1993); and W.J. STANKIEWICZ, *In Search of a Political Philosophy: Ideologies at the Close of the Twentieth Century* (1993).

Socialism. DANIEL BELL, "Socialism," in *International Encyclopedia of the Social Sciences,* 14:506–534 (1968), is one of the best short surveys of socialism from its origins to the mid-1950s. G.D.H. COLE, *A History of Socialist Thought,* 5 vol. in 7 (1953–60), provides a detailed general study. Two histories written from a social democratic perspective are HARRY W. LAIDLER, *History of Socialism,* updated and expanded ed. (1968); and CARL LANDAUER, ELIZABETH KRIDL VALKENIER, and HILDE STEIN LANDAUER, *European Socialism,* 2 vol. (1959, reprinted 1976). FRITZ U. SCHARPF, *Crisis and Choice in European Social*

Democracy (1991; originally published in German, 1987), analyzes the status of social democracy in Europe to the mid-1980s. RADOSLAV SELUCKÝ, *Marxism, Socialism, Freedom* (1979), is also useful. GORDON WHITE, ROBIN MURRAY, and CHRISTINE WHITE (eds.), *Revolutionary Socialist Development in the Third World* (1983); and MELANIE A. SULLY, *Continuity and Change in Austrian Socialism* (1982), are two good resources on socialism as distinct from social democracy. IRVING HOWE (ed.), *Essential Works of Socialism*, 3rd ed. (1986), collects readings in many varieties of Marxist thought. MICHAEL HARRINGTON, *Why We Need Socialism in America* (1970), is a defense of the relevance of socialism in the United States by a prominent American socialist.

JAMES JOLL, *The Second International, 1889–1914*, rev. and extended ed. (1974), offers a standard but very readable history of 19th-century social democracy. GEORGE LICHTHEIM, *Marxism*, 2nd ed. rev. (1964, reprinted 1982), a sophisticated and searching work, deals with the transformation of Marxism from the time of its founders to that of Lenin and Stalin, while his *A Short History of Socialism* (1970, reissued 1983), and *The Origins of Socialism* (1969), are idiosyncratic but stimulating reading. ADOLF STURMTHAL, *The Tragedy of European Labor* (1943), is a study of the failures of social democratic policies during the depression of the 1930s, by a sympathetic observer.

Additional studies include BRANKO HORVAT, *The Political Economy of Socialism* (1982); ANTHONY GIDDENS, *A Contemporary Critique of Historical Materialism*, 2 vol. (1982–85); TOM BOTTOMORE, *Sociology and Socialism* (1984); and MICHAEL ELLMAN, *Socialist Planning*, 2nd ed. (1989).

Communism. Useful works include SHLOMO AVINERI, *The Social and Political Thought of Karl Marx* (1968, reissued 1990; originally published in Hebrew, 1967); TOM BOTTOMORE (ed.), *Interpretations of Marx* (1988); LESZEK KOŁAKOWSKI, *Main Currents of Marxism*, 3 vol. (1978; originally published in Polish, 1976–78); H.B. ACTON, *The Illusion of the Epoch* (1962, reissued 1972); ROBERT C. TUCKER, *Philosophy and Myth in Karl Marx*, 2nd ed. (1972); ISAIAH BERLIN, *Karl Marx: His Life and Environment*, 4th ed. (1978); R.N. CAREW HUNT, *The Theory and Practice of Communism*, 5th rev. ed. (1957, reissued 1977); J.L.H. KEEP, *The Rise of Social Democracy in Russia* (1963), an outstanding history of the subject up to 1906; LEONARD SCHAPIRO, *The Origin of the Communist Autocracy: Political Opposition in the Soviet State: First Phase, 1917–1922*, 2nd ed. (1977); ADAM B. ULAM, *The Bolsheviks* (1965; also published as *Lenin and the Bolsheviks*, 1966); FRANCO VENTURI, *Roots of Revolution: A History of the Populist and Socialist Movements in Nineteenth Century Russia* (1960, reprinted 1983; originally published in Italian, 1952); and BERTRAM D. WOLFE, *Three Who Made a Revolution*, 4th rev. ed. (1964, reissued 1984), a readable, stimulating history of Bolshevism in its formative years.

Works on Stalinism include IAN GREY, *The First Fifty Years: Soviet Russia, 1917–67* (1967); LEONARD SCHAPIRO, *The Communist Party of the Soviet Union*, 2nd rev. and enlarged ed. (1970), a detailed history of the Communist Party in theory and in practice up to 1968; ROBERT CONQUEST, *The Great Terror: Stalin's Purge of the Thirties*, rev. ed. (1973); IVO BANAC, *With Stalin Against Tito* (1988); GRAEME GILL, *Stalinism* (1990); and ROBERT V. DANIELS, *Trotsky, Stalin, and Socialism* (1991).

The phenomenon of totalitarianism is treated well in HANNAH ARENDT, *The Burden of Our Time* (1951; also published as *The Origins of Totalitarianism*, new ed., 1973, reprinted 1986); and ELLEN FRANKEL PAUL (ed.), *Totalitarianism at the Crossroads* (1990).

The world movement up to Stalin's death is treated in HAMILTON FISH ARMSTRONG, *Tito and Goliath* (1951), an excellent study of the conflict between Yugoslavia and the Soviet Union; ZBIGNIEW K. BRZEZINSKI, *The Soviet Bloc: Unity and Conflict*, rev. and enlarged ed. (1967), and *Out of Control: Global Turmoil on the Eve of the Twenty-First Century* (1993); CONRAD BRANDT, BENJAMIN SCHWARTZ, and JOHN K. FAIRBANK, *A Documentary History of Chinese Communism* (1952, reissued 1971); LIU KANG and XIAOBING TANG (eds.), *Politics, Ideology, and Literary Discourse in Modern China* (1993); VLADIMIR DEDIJER, *Tito Speaks* (1953); and JANE DEGRAS (ed.), *Communist International, 1919–1943: Documents*, 3 vol. (1956–65, reprinted 1971), a collection of documents with commentary.

Also of interest are HERBERT FEIS, *Churchill, Roosevelt, Stalin*, 2nd ed. (1967), a well-documented study of wartime diplomacy; GUNTHER NOLLAU, *International Communism and World Revolution* (1961, reissued 1975; originally published in German, 1959); EUGENIO REALE, *Avec Jacques Duclos: au banc des accusés à la réunion constitutive du Kominform à Szklarska Poreba (22–27 septembre 1947)*, trans. from Italian (1958); DAVID REES, *Korea: The Limited War* (1964); and HUGH SETON-WATSON, *From Lenin to Khrushchev* (1960, reissued 1985; also published as *The Pattern of Communist Revolution*, rev. and enlarged ed., 1960), a study of the rise of Communism in eastern Europe.

Developments after Stalin are surveyed in TARIQ ALI (ed.), *The Stalinist Legacy* (1984); ADAM BROMKE (ed.), *The Communist States at the Crossroads Between Moscow and Peking* (1965); ALEXANDER DALLIN (ed.), *Diversity in International Communism: A Documentary Record, 1961–63* (1963); HÉLÈNE CARRÈRE D'ENCAUSSE and STUART R. SCHRAM, *Marxism and Asia: An Introduction with Readings* (1969; originally published in French, 1965); EDWARD CRANKSHAW, *The New Cold War: Moscow v. Peking* (1963); GHIŢA IONESCU, *The Break-up of the Soviet Empire in Eastern Europe* (1965, reprinted 1984); WALTER LAQUEUR and LEOPOLD LABEDZ (eds.), *Polycentrism* (1962), a valuable collection of essays on dissent in the Communist parties; WOLFGANG LEONHARD, *The Kremlin Since Stalin* (1962, reprinted 1975; originally published in German, 1959); and COLUMBIA UNIVERSITY, RUSSIAN INSTITUTE, *The Anti-Stalin Campaign and International Communism*, rev. ed. (1956), an annotated text of Khrushchev's secret speech in 1956, with some other documents.

Additional studies include HUGH SETON-WATSON, *The Imperialist Revolutionaries* (1978); H. GORDON SKILLING, *The Governments of Communist East Europe* (1966); MICHEL TATU, *Power in the Kremlin* (1969; originally published in French, 1967); DONALD S. ZAGORIA, *The Sino-Soviet Conflict, 1956–61* (1962, reissued 1973); MILORAD M. DRACHKOVITCH (ed.), *Fifty Years of Communism in Russia* (1968), a symposium by a number of experts on changes in Communist doctrine; LEONARD SCHAPIRO (ed.), *The USSR and the Future* (1962), essays by specialists on various aspects of the party program of 1961; PETER FERDINAND, *Communist Regimes in Comparative Perspective* (1991); and GALE STOKES, *The Walls Came Tumbling Down* (1993).

Three important sources for studies of regime change are DANIEL CHIROT, *Social Change in the Modern Era*, ed. by ROBERT K. MERTON (1986); JACK A. GOLDSTONE (ed.), *Revolutions: Theoretical, Comparative, and Historical Studies*, 2nd ed. (1994); and JOHN DUNN, *Modern Revolutions*, 2nd ed. (1989).

Anarchism. *General:* The best general accounts of anarchism are PETER MARSHALL, *Demanding the Impossible: A History of Anarchism* (1992); JAMES JOLL, *The Anarchists*, 2nd ed. (1980); PAUL AVRICH, *Anarchist Portraits* (1988); GEORGE WOODCOCK, *Anarchism*, new ed. (1986); HAROLD BARCLAY, *People Without Government*, completely rev. ed. (1990); DANIEL GUÉRIN, *Anarchism: From Theory to Practice* (1970; originally published in French, 1965); PAUL ELTZBACHER, *Anarchism* (1960, reprinted 1972; originally published in German, 1900); and RICHARD D. SONN, *Anarchism* (1992).

Good anthologies of anarchist theory include IRVING LOUIS HOROWITZ (ed.), *The Anarchists* (1964); LEONARD I. KRIMERMAN and LEWIS PERRY (eds.), *Patterns of Anarchy* (1966); and DAVID E. APTER and JAMES JOLL (eds.), *Anarchism Today* (1971).

For a comprehensive bibliography of anarchist literature from different countries over the last two centuries, see DENISE FAUVEL-ROUIF (ed.), *Anarchism* (1982); and HELENE STRUB (ed.), *Anarchism*, vol. 2 (1993). There are also useful selected bibliographies in all the books listed above.

Classic authors: The earliest formulations of modern anarchist thought can be found in WILLIAM GODWIN, *An Enquiry Concerning Political Justice and Its Influence on General Virtue and Happiness*, 2 vol. (1792). PIERRE-JOSEPH PROUDHON's key anarchist work is *Qu'est-ce que la propriété?* (1840; *What Is Property?*, trans. by BENJAMIN R. TUCKER, 1876). Of PETER KROPOTKIN's many writings, his *Memoirs of a Revolutionist* (1899) is essential. See also his *La Conquête du pain* (1892; *The Conquest of Bread*, 1906), *Fields, Factories, and Workshops* (1899), and *Mutual Aid: A Factor of Evolution* (1902).

LEO TOLSTOY's Christian anarchist concepts can be found in many of his later works, including *T Sarstvo Bozhie vnutri nas* (1894; *The Kingdom of God Is Within You*, trans. by CONSTANCE GARNETT, 1894), and *V chem moîa vera* (1884; *What I Believe*, trans. by CONSTANTINE POPOFF, 1885). EMMA GOLDMAN's voluminous writings include her autobiography, *Living My Life*, 2 vol. (1931, reissued 1988); see also the autobiography of Goldman's comrade ALEXANDER BERKMAN, *Prison Memoirs of an Anarchist* (1912, reprinted 1970).

For an analysis of Godwin's work, see ISAAC KRAMNICK, *The Politics of Political Philosophy, A Case Study: Godwin's Anarchism and Radical England* (1970). For information on Proudhon's life and thought, see STEWART EDWARDS (ed.), *Selected Writings of Pierre-Joseph Proudhon* (1969); K. STEVEN VINCENT, *Pierre-Joseph Proudhon and the Rise of French Republican Socialism* (1984); ALAN RITTER, *The Political Thought of Pierre-Joseph Proudhon* (1969, reprinted 1980); and ROBERT L. HOFFMAN, *Revolutionary Justice: The Social and Political Theory of P.-J. Proudhon* (1972).

For an anthology of Kropotkin's writings, see MARTIN A. MILLER (ed.), *Peter Kropotkin: Selected Writings on Anarchism*

and Revolution (1970). On Kropotkin's life, consult MARTIN A. MILLER, *Kropotkin* (1976); CAROLINE CAHM, *Kropotkin and the Rise of Revolutionary Anarchism, 1872–1886* (1989); and GEORGE WOODCOCK and IVAN AVAKUMOVIC, *The Anarchist Prince: A Biographical Study of Peter Kropotkin* (1950, reissued 1970). Two convenient volumes that explore the anarchist ideas of Michael Bakunin are ARTHUR LEHNING (ed.), *Selected Writings [of] Michael Bakunin* (1973); and SAM DOLGOFF (ed.), *Bakunin on Anarchism*, 2nd rev. ed. (1980; originally published as *Bakunin on Anarchy*, 1972). For biographies of Bakunin, see EDWARD H. CARR, *Michael Bakunin* (1937, reissued 1975); ARTHUR P. MENDEL, *Michael Bakunin: Roots of Apocalypse* (1981); and EUGENE PYZIUR, *The Doctrine of Anarchism of Michael A. Bakunin* (1955, reissued 1968).

Articles by EMMA GOLDMAN are collected in her *Anarchism and Other Essays*, 3rd rev. ed. (1917, reprinted 1967); and in ALIX KATES SHULMAN (ed.), *Red Emma Speaks*, 3rd ed. (1996). For a sympathetic biography of Goldman, see RICHARD DRINNON, *Rebel in Paradise* (1961, reissued 1982). The correspondence of Goldman and Berkman can be found in RICHARD DRINNON and ANNA MARIA DRINNON (eds.), *Nowhere at Home* (1975). The life of Johann Most is studied in FREDERIC TRAUTMANN, *The Voice of Terror: A Biography of Johann Most* (1980).

Russian anarchist thought: The best general account of the anarchist movement in Russia remains PAUL AVRICH, *The Russian Anarchists* (1967, reprinted 1980). For documents on the anarchist critique of Lenin and bolshevism, see PAUL AVRICH (ed.), *Anarchists in the Russian Revolution* (1973). VOLINE, *The Unknown Revolution, 1917–1921* (1955; originally published in French, 1947), is a good anarchist memoir of the Russian Revolution.

Anarchism in the United States: On the origins and history of American individualist anarchism, see CARLOTTA R. ANDERSON, *All-American Anarchist: Joseph A. Labadie and the Labor Movement* (1998); JAMES J. MARTIN, *Men Against the State: The Expositors of Individualist Anarchism in America, 1827–1908* (1953, reissued 1970); and RUDOLF ROCKER, *Pioneers of American Freedom*, trans. from German (1949). An excellent collection of historical source material can be found in PAUL AVRICH, *Anarchist Voices: An Oral History of Anarchism in America* (1995).

There is a large literature on Haymarket. Especially useful are PAUL AVRICH, *The Haymarket Tragedy* (1984); and DAVID ROEDIGER and FRANKLIN ROSEMONT (eds.), *Haymarket Scrapbook* (1986). Anarchist influences in the Industrial Workers of the World are discussed in SALVATORE SALERNO, *Red November, Black November: Culture and Community in the Industrial Workers of the World* (1989).

Anarchism in Europe: P. HOLGATE, *Malatesta* (1956), is a study of the leading Italian anarchist. See also VERNON RICHARDS (ed.), *Errico Malatesta: His Life & Ideas*, 3rd ed. (1984). On France, see MARIE FLEMING, *The Anarchist Way to Socialism: Elisée Reclus and Nineteenth-Century European Anarchism* (1979). On Germany, the best book is ANDREW R. CARLSON, *Anarchism in Germany* (1972). See also EUGENE LUNN, *Prophet of Community: The Romantic Socialism of Gustav Landauer* (1973). On Spain, see TEMMA KAPLAN, *Anarchists of Andalusia* (1977); and ROBERT W. KERN, *Red Years/Black Years: A Political History of Spanish Anarchism, 1911–1937* (1978). For developments in England, see HERMIA OLIVER, *The International Anarchist Movement in Late Victorian London* (1983).

Contemporary anarchism: DANIEL COHN-BENDIT and GABRIEL COHN-BENDIT, *Obsolete Communism* (1968; originally published in French, 1968), by two participants in the 1968 student uprisings in Paris, effectively describes anarchist involvement in the protests; its critique of the contemporary French Communist Party is richly informed by a historical analysis of early anarchist resistance to bolshevism in Russia. DAVID APTER and JAMES JOLL (eds.), *Anarchism Today* (1971), is a very good summation of the influence of anarchism around the world in the aftermath of the student uprisings of the 1960s. NOAM CHOMSKY, *Radical Priorities*, 2nd rev. ed., edited by CARLOS P. OTERO (1984), includes a good sample of Chomsky's anarchism-related writings.

MARGARET S. MARSH, *Anarchist Women, 1870–1920* (1981), treats the early development of feminist anarchism. Essays on contemporary anarcha-feminism by ELAINE LEEDER, SUSAN BROWN, PEGGY KORNEGGER, and CAROL ERLICH appear in HOWARD J. EHRLICH (ed.), *Reinventing Anarchy, Again* (1996).

The most influential thinking in contemporary anarchism can be found in the work of MURRAY BOOKCHIN, *Post-Scarcity Anarchism*, 2nd ed. (1986), and *The Modern Crisis*, 2nd rev. ed. (1987).

Fascism. *General studies:* STANLEY G. PAYNE, *A History of Fascism, 1914–1945* (1995), discusses fascist and extreme right movements in several countries, including the United States. Analyses of fascism in various European countries are presented in ALEXANDER DE GRAND, *Fascist Italy and Nazi Germany* (1995); S.J. WOOLF (ed.), *European Fascism* (1968); JOHN WEISS, *The Fascist Tradition: Radical Right-Wing Extremism in Modern Europe* (1967); and WALTER LAQUEUR and GEORGE L. MOSSE (eds.), *International Fascism, 1920–1945* (1966). ARNO J. MAYER, *Dynamics of Counterrevolution in Europe, 1870–1956: An Analytic Framework* (1971), is a study of collaboration and divergence between counterrevolutionaries and fascists. ALASTAIR HAMILTON, *The Appeal of Fascism: A Study of Intellectuals and Fascism, 1919–1945* (1971), discusses fascism in Italy, Germany, France, and Britain.

On the social bases of European fascism, see STEIN UGELVIK LARSEN, BERTN HAGTVET, and JAN PETTER MYKLEBUST (eds.), *Who Were the Fascists: Social Roots of European Fascism* (1980); DETLEF MÜHLBERGER (ed.), *The Social Basis of European Fascist Movements* (1987); and CHARLES S. MAIER, *Recasting Bourgeois Europe: Stabilization in France, Germany, and Italy in the Decade After World War I* (1975, reprinted 1988). ROGER GRIFFEN (ed.), *International Fascism* (1998), covers various theoretical approaches to fascism. Also of interest is ROBERT O. PAXTON, "The Five Stages of Fascism," *Journal of Modern History*, 70(1) (March 1998), pp. 1–23; and ROBERT J. SOUCY, "Functional Hating: French Fascist Demonology Between the Wars," in *Contemporary French Civilization*, 23(2) (Summer/Fall 1999), pp. 158–176.

Italy: Basic works on Italian fascism include ALEXANDER DE GRAND, *Italian Fascism: Its Origins & Development*, 3rd ed. (2000); RENZO DE FELICE, *Fascism: An Informal Introduction to Its Theory and Practice* (1977); and EDWARD R. TANNENBAUM, *The Fascist Experience: Italian Society and Culture, 1922–1945* (1972). The experience of Jews in Italy is covered in SUSAN ZUCCOTTI, *The Italians and the Holocaust: Persecution, Rescue, and Survival* (1987, reissued 1996); and MICHAEL LEDEEN, "Italian Jews and Fascism," *Judaism*, 18(3) (Summer 1969), pp. 277–298. The role of women is discussed in VICTORIA DE GRAZIA, *How Fascism Ruled Women: Italy, 1922–1945* (1992).

Germany: Analyses of the social bases of Nazism include SHELLEY BARANOWSKI, *The Sanctity of Rural Life: Nobility, Protestantism, and Nazism in Weimar Prussia* (1995), and *The Confessing Church, Conservative Elites, and the Nazi State* (1986); MICHAEL H. KATER, *The Nazi Party: A Social Profile of Members and Leaders, 1919–1945* (1983); and, on the Nazi Christian movement, DORIS L. BERGEN, *Twisted Cross: The German Christian Movement in the Third Reich* (1996). Intellectual and cultural precursors are covered in GEORGE L. MOSSE, *Nazi Culture: Intellectual, Cultural, and Social Life in the Third Reich* (1966), and *The Crisis of German Ideology: Intellectual Origins of the Third Reich* (1964, reprinted 1981). Also of interest are IAN KERSHAW, *Popular Opinion and Political Dissent in the Third Reich, Bavaria, 1933–1945* (1983); and DAVID SCHOENBAUM, *Hitler's Social Revolution* (1966, reprinted 1980).

France, Russia, and Spain: Important studies of French fascism include ROBERT J. SOUCY, *French Fascism: The Second Wave, 1933–1939* (1995), *French Fascism: The First Wave, 1924–1933* (1986), and *Fascist Intellectual: Drieu La Rochelle* (1979); ZEEV STERNHELL, *Neither Right nor Left: Fascist Ideology in France*, trans. by DAVID MAISEL (1986); and EUGEN JOSEPH WEBER, *Action Française: Royalism and Reaction in Twentieth Century France* (1962). Fascist movements in Russia and Spain are discussed in WALTER LAQUEUR, *Black Hundred: The Rise of the Extreme Right in Russia* (1993); and STANLEY G. PAYNE, *Falange: A History of Spanish Fascism* (1961).

Fascism outside Europe: Various non-European movements are covered in STANLEY G. PAYNE, *A History of Fascism, 1914–1945*, cited above. Historical and contemporary studies of Argentine fascism are DAVID ROCK, *Authoritarian Argentina: The Nationalist Movement, Its History, and Its Impact* (1993); and DANIEL JAMES, *Resistance and Integration: Peronism and the Argentine Working Class, 1946–1976* (1988, reissued 1993). JOHN DIGGINS, *Mussolini and Fascism: The View from America* (1972), discusses support for Italian fascism in the United States. Japanese fascism is covered in MASAO MARUYAMA, *Thought and Behavior in Modern Japanese Politics*, expanded ed., edited by IVAN MORRIS (1969).

Neofascism: RICHARD GOLSAN, *Fascism's Return* (1998), is a broad survey. Neofascism in Italy, Germany, France, and Britain is discussed in ROGER EATWELL, *Fascism: A History* (1995). LUCIANO CHELES, RONNIE FERGUSON, and MICHALINA VAUGHN (eds.), *The Far Right in Western and Eastern Europe* (1995), covers several neofascist movements. French neofascism is discussed in JIM WOLFREYS, "Neither Right nor Left? Towards an Integrated Analysis of the Front National," in NICHOLAS ATAKIN and FRANK TALLET (eds.), *The Right in France, 1789–1997* (1998). Russian neofascism is covered in SVEN GUNNAR SIMONSEN, *Politics and Personalities: Key Actors in the Russian Opposition* (1996); and WALTER LAQUEUR, *Black Hundred: The Rise of the*

Extreme Right in Russia, cited above. JACQUES JULLIARD, *Ce fascisme qui vient* (1994), is a study of neofascism in the former Yugoslavia. Chilean neofascism is discussed in MARY HELEN SPOONER, *Soldiers in a Narrow Land: The Pinochet Regime in Chile* (1994).

Nationalism. The origins and development of nationalism as a political idea are discussed in detail in E.J. HOBSBAWM, *Nations and Nationalism Since 1780*, 2nd ed. (1992); ELIE KEDOURIE, *Nationalism*, 4th, expanded ed. (1993); LOUIS L. SNYDER, *Encyclopedia of Nationalism* (1990); ERNEST GELLNER, *Nations and Nationalism* (1983); and HANS KOHN, *The Idea of Nationalism* (1944, reissued 1967). Other discussions of the development of nationalism, including various interpretations in modern times, are BENEDICT ANDERSON, *Imagined Communities*, rev. and extended ed. (1991); LIAH GREENFELD, *Nationalism* (1992); JOHN LUKACS, *The End of the Twentieth Century and the End of the Modern Age* (1993); and LOUIS L. SNYDER, *Varieties of Nationalism* (1976). A sociological treatment of the subject is ANTHONY D. SMITH, *Theories of Nationalism*, 2nd ed. (1983). HUGH SETON-WATSON, *Nations and States* (1977), is a detailed, worldwide study.

A new approach to the problem of nationalism was introduced by KARL W. DEUTSCH, *Nationalism and Social Communication*, 2nd ed. (1965), and *Tides Among Nations* (1979), an overview of the development of his thinking. Nationalism outside western Europe is discussed in GREGORY MADDOX (ed.), *African Nationalism and Revolution* (1993); DAVID ROCK, *Authoritarian Argentina: The Nationalist Movement, Its History, and Its Impact* (1993); JILL A. IRVINE, *The Croat Question: Partisan Politics in the Formation of the Yugoslav Socialist State* (1993); RUPERT EMERSON, *From Empire to Nation: The Rise to Self-Assertion of Asian and African Peoples* (1960, reissued 1970); ELIE KEDOURIE (ed.), *Nationalism in Asia and Africa* (1970); and SELIG S. HARRISON, *The Widening Gulf* (1978).

There are two valuable though dated bibliographies: KOPPEL S. PINSON, *A Bibliographical Introduction to Nationalism* (1935); and KARL W. DEUTSCH, *Interdisciplinary Bibliography on Nationalism* (1956).

Liberalism. General works include CAROLE PATEMAN, *Participation and Democratic Theory* (1970); MAX WEBER, *Economy and Society*, ed. by GUENTHER ROTH and CLAUS WITTICH, 3 vol. (1968; originally published in German, 4th rev. ed., 1956); ÉMILE DURKHEIM, *Durkheim on Politics and the State*, ed. by ANTHONY GIDDENS, trans. from German (1986); ANTHONY GIDDENS, *Capitalism and Modern Social Theory* (1971); PATRICK DUNLEAVY and BRENDAN O'LEARY, *Theories of the State: The Politics of Liberal Democracy* (1987); LARRY DIAMOND and MARC F. PLATTNER (eds.), *Capitalism, Socialism, and Democracy Revisited* (1991); ROBERT DENOON CUMMING, *Human Nature and History: A Study of the Development of Liberal Political Thought* (1969); and HARRY K. GIRVETZ, *Evolution of Liberalism*, rev. ed. (1963), a comparison of classical and contemporary liberalism that examines the psychological assumptions underlying the political and economic views of the classical liberals.

Also of interest are LOUIS HARTZ, *The Liberal Tradition in America: An Interpretation of American Political Thought Since the Revolution* (1955, reprinted 1991); KINGSLEY MARTIN, *French Liberal Thought in the Eighteenth Century*, 2nd ed., rev. (1954, reprinted as *The Rise of French Liberal Thought*, 1980); K.R. MINOGUE, *The Liberal Mind* (1963), on the contention that liberalism reflects a moral and political consensus; and FREDERICK WATKINS, *The Political Tradition of the West: A Study in the Development of Modern Liberalism* (1948, reprinted 1982).

Classical liberalism is treated in A.V. DICEY, *Lectures on the Relation Between Law and Public Opinion in England During the Nineteenth Century*, 2nd ed. (1914, reprinted 1981); E. HALÉVY, *The Growth of Philosophical Radicalism* (1928, reissued 1972; originally published in French, 3 vol., 1901–04); L.T. HOBHOUSE, *Liberalism* (1911, reissued 1980); HAROLD J. LASKI, *The Rise of Liberalism: The Philosophy of a Business Civilization* (1936; also published as *The Rise of European Liberalism*, 1936, reissued 1971); GUIDO DE RUGGIERO, *The History of European Liberalism* (1927, reissued 1981; originally published in Italian, 1925); and LESLIE STEPHEN, *The English Utilitarians*, 3 vol. (1900, reprinted 1968).

Works on contemporary liberalism include ADOLF A. BERLE and GARDINER C. MEANS, *The Modern Corporation and Private Property* (1933, reissued 1991), the classic account of how the private corporation has generated managerial power; JOHN MAYNARD KEYNES, *The End of Laissez-faire* (1926), and *The General Theory of Employment, Interest, and Money* (1935, reissued 1991), the most influential economic tract of the first half of the 20th century; WILLIAM H. BEVERIDGE, *Full Employment in a Free Society*, 2nd ed. (1960), by the chief English architect of the welfare state; G.D.H. COLE, *Economic Planning* (1935, reissued 1971); JOHN DEWEY, *Human Nature and Conduct: An Introduction to Social Psychology* (1922, reissued 1957), the best criticism of the psychological preconceptions of the classical liberals; JOHN KENNETH GALBRAITH, *American Capitalism: The Concept of Countervailing Power*, rev. ed. (1956, reissued 1993), *The Affluent Society*, 4th ed. (1984), one of the first works to contend that liberal emphasis should shift from greater productivity to the question of priorities, and *The New Industrial State*, 4th ed. (1985); ALVIN H. HANSEN, *Economic Policy and Full Employment* (1947), by a prominent American exponent of Keynesianism; STEPHEN HOLMES, *The Anatomy of Antiliberalism* (1993); JOHN GRAY, *Beyond the New Right: Markets, Government, and the Common Environment* (1993); and SEYMOUR E. HARRIS, *Economic Planning* (1949), an examination of economic planning in 10 European countries as well as in the United States, Argentina, Japan, and India.

Additional studies include W. ARTHUR LEWIS, *The Principles of Economic Planning*, 3rd ed. (1969), on the difference between direct and indirect planning and its important implications; JAMES EDWARD MEADE, *Planning and the Price Mechanism* (1949), a liberal socialist approach to the problems of planning and use of the price mechanism in postwar Britain; GUNNAR MYRDAL, *Beyond the Welfare State* (1960, reprinted 1982), on the social impact of economic planning in highly industrialized states and the economic relationships between them and less developed economies; STEVEN M. DELUE, *Political Obligation in a Liberal State* (1989), on the social impact of post-Keynesianism; EUGENE V. ROSTOW, *Planning for Freedom* (1959), an analysis of the impact of the Keynesian "revolution" on government intervention in the economic sphere; and ANDREW LEVINE, *Liberal Democracy: A Critique of its Theory* (1981).

Conservatism. Useful anthologies include PETER VIERECK, *Conservatism: From John Adams to Churchill* (1956, reprinted 1978); and PETER WITONSKI (ed.), *The Wisdom of Conservatism* (1971).

Among Burkean-conservative works sharing an anti-extremist centre with moderate liberals are DANIEL J. BOORSTIN, *The Genius of American Politics* (1953, reissued 1966); and THOMAS I. COOK and MALCOLM MOOS, *Power Through Purpose* (1954). Other studies include ROSS J.S. HOFFMAN and PAUL LEVACK (eds.), *Burke's Politics* (1949, reissued 1967); HENRY A. KISSINGER, *A World Restored* (1957, reissued 1973); WALTER LIPPMANN, *The Cold War* (1947, reissued 1972); REINHOLD NIEBUHR, *The Irony of American History* (1952, reissued 1984); ROBERT A. NISBET, *Community and Power* (1962); LEO STRAUSS, *Natural Right and History* (1953, reissued 1971), and *What Is Political Philosophy? and Other Studies* (1959, reprinted 1988); PETER VIERECK, *Conservatism Revisited*, rev. and enlarged ed. (1965, reprinted 1978); ERIC VOEGELIN, *Order and History*, 5 vol. (1956–87); FRANCIS GRAHAM WILSON, *The Case for Conservatism* (1951, reissued 1990); and ROGER SCRUTON, *The Meaning of Conservatism* (1980).

Conservatism in Britain is discussed in L.S. AMERY, *The Forward View* (1935, reissued 1971); ARTHUR BRYANT, *The Spirit of Conservatism* (1929); HUGH CECIL, *Conservatism* (1912, reissued 1937); HENRY FAIRLIE, *The Life of Politics* (1968); QUINTIN HOGG, *The Conservative Case*, rev. ed. (1959); F.J.C. HEARNSHAW, *Conservatism in England* (1933, reprinted 1968); MICHAEL OAKESHOTT, *Rationalism in Politics, and Other Essays*, new and expanded ed. (1991); JAMES J. SACK, *From Jacobite to Conservative: Reaction and Orthodoxy in Britain, c. 1760–1832* (1993); and PEREGRINE WORSTHORNE, *The Socialist Myth* (1971).

Other countries are treated in LARRY EUGENE JONES and JAMES RETALLACK (eds.), *Between Reform, Reaction, and Resistance: Studies in the History of German Conservatism from 1789 to 1945* (1993); ARTHUR AUGHEY, GRETA JONES, and W.T.M. RICHES, *The Conservative Political Tradition in Britain and the United States* (1992); T.J. PEMPEL, *Policy and Politics in Japan: Creative Conservatism* (1982); MARTIN E. MARTY and R. SCOTT APPLEBY (eds.), *Fundamentalisms Observed* (1991), a worldwide study; KATHARINE WEST, *The Revolution in Australian Politics* (1984); and PAUL GIFFORD, *The New Crusaders: Christianity and the New Right in Southern Africa* (1991).

BURTON YALE PINES, *Back to Basics* (1982), recounts the conservative resurgence of the 1970s in the United States. Perspectives and overviews linked to the resurgence of conservatism in the 1980s and to post-Keynesianism can be found in PETER H. MERKL and LEONARD WEINBERG (eds.), *Encounters with the Contemporary Radical Right* (1993); BRUCE FROHNEN, *Virtue and the Promise of Conservatism* (1993); and PAUL HAINSWORTH (ed.), *The Extreme Right in Europe and the USA* (1992).

A variety of more radical conservative views can be found in WILLIAM F. BUCKLEY, *God and Man at Yale: The Superstitions of Academic Freedom* (1951, reissued 1986), and *Up from Liberalism*, rev. ed. (1968); JAMES BURNHAM, *Congress and the American Tradition* (1959), and *Suicide of the West* (1964, reissued 1985); MILTON FRIEDMAN, *Capitalism and Freedom* (1962, reis-

sued 1982); BARRY GOLDWATER, *Conscience of a Conservative* (1960, reissued 1990); JEFFREY HART, *The American Dissent* (1966); NELLIE D. KENDALL (ed.), *Willmoore Kendall contra mundum* (1971); RUSSELL KIRK, *The Conservative Mind: From Burke to Eliot*, 7th rev. ed. (1986), *Prospects for Conservatives* (1956, reissued 1989), and *Enemies of the Permanent Things*, rev. ed. (1984); ERIK VON KUEHNELT-LEDDIHN, *Liberty or Equality* (1952, reissued 1993); FRANK S. MEYER, *In Defense of Freedom* (1962), and *The Conservative Mainstream* (1969); THOMAS MOLNAR, *The Counter-Revolution* (1969); J. ENOCH POWELL, *Freedom and Reality* (1969); and RONALD REAGAN, *The Creative Society* (1968). MICHAEL LIENESCH, *Redeeming America: Piety and Politics in the New Christian Right* (1993), discusses conservatism and religion.

(L.A.C./G.W./Ar.Di./Fr.R./M.M./Ro.So./H.K./H.K.G./Pe.V./Ed.)

Socrates

Socrates of Athens, who flourished in the last half of the 5th century BC, was the first of the great trio of ancient Greeks—Socrates, Plato, and Aristotle—who laid the philosophical foundations of Western culture. His profound influence on ancient and modern Western philosophy stems not only from his thought but also from his way of life and his character.

Socrates was a widely recognized and controversial figure in his native Athens, so much so that he was frequently mocked in the plays of comic dramatists. (The *Clouds* of Aristophanes, produced in 423, is the best-known example.) Although Socrates himself wrote nothing, he is depicted in conversation in compositions by a small circle of his admirers—Plato and Xenophon first among them. He is portrayed in these works as a man of great insight, integrity, self-mastery, and argumentative skill. The impact of his life was all the greater because of the way in which it ended: at age 70, he was brought to trial on a charge of impiety and sentenced to death by poisoning (the poison probably being hemlock) by a jury of his fellow citizens. Plato's *Apology of Socrates* purports to be the speech Socrates gave at his trial in response to the accusations made against him (Greek *apologia* means "defense"). Its powerful advocacy of the examined life and its condemnation of Athenian democracy have made it one of the central documents of Western thought and culture.

The article is divided into the following sections:

Philosophical and literary sources 438
- Xenophon
- Plato
- Aristotle

Life and personality 440

Background of the trial 440A
- Religious scandal and the coup of the oligarchs
- The perceived fragility of Athenian democracy
- The Athenian ideal of free speech

Plato's *Apology* 440C

The public's hatred of Socrates 440D
- The impression created by Aristophanes
- The human resistance to self-reflection
- Socrates' criticism of democracy

The charge of impiety 440E
- Socrates' radical reconception of piety
- The danger posed by Socrates

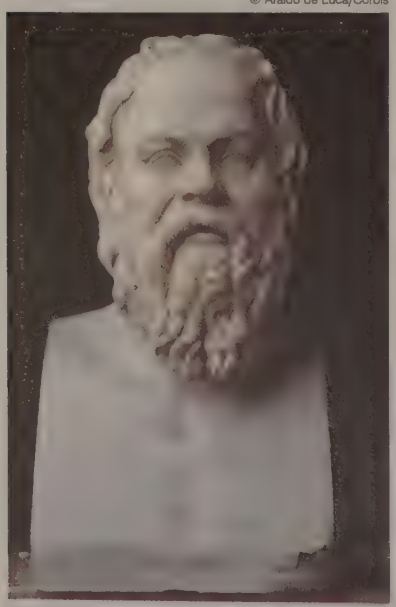
Socrates versus Plato 440F

The legacy of Socrates 440F

Bibliography 440G

Philosophical and literary sources

While Socrates was alive, he was, as noted, the object of comic ridicule, but most of the plays that make reference to him are entirely lost or exist only in fragmentary form—*Clouds* being the chief exception. Although Socrates is the central figure of this play, it was not Aristophanes' purpose to give a balanced and accurate portrait of him (comedy never aspires to this) but rather to use him to represent certain intellectual trends in contemporary Athens—the study of language and nature and, as Aristophanes implies, the amoralism and atheism that accompany these pursuits. The value of the play as a reliable source of knowledge about Socrates is thrown further into doubt by the fact that, in Plato's *Apology*, Socrates himself rejects it as a fabrication. This aspect of the trial will be discussed more fully below.



Socrates, herm from a Greek original, second half of the 4th century BC; in the Capitoline Museums, Rome.

Soon after Socrates' death, several members of his circle preserved and praised his memory by writing works that represent him in his most characteristic activity—conversation. His interlocutors in these (typically adversarial) exchanges included people he happened to meet, devoted followers, prominent political figures, and leading thinkers of the day. Many of these "Socratic discourses," as Aristotle calls them in his *Poetics*, are no longer extant; there are only brief remnants of the conversations written by Antisthenes, Aeschines, Phaedo, and Eucleides. But those composed by Plato and Xenophon survive in their entirety. What knowledge we have of Socrates must therefore depend primarily on one or the other (or both, when their portraits coincide) of these sources. (Plato and Xenophon also wrote separate accounts, each entitled *Apology of Socrates*, of Socrates' trial.) Most scholars, however, do not believe that every Socratic discourse of Xenophon and Plato was intended as a historical report of what the real Socrates said, word-for-word, on some occasion. What can reasonably be claimed about at least some of these dialogues is that they convey the gist of the questions Socrates asked, the ways in which he typically responded to the answers he received, and the general philosophical orientation that emerged from these conversations.

XENOPHON

Among the compositions of Xenophon, the one that gives the fullest portrait of Socrates is *Memorabilia*. The first two chapters of Book I of this work are especially important, because they explicitly undertake a refutation of the charges made against Socrates at his trial; they are therefore a valuable supplement to Xenophon's *Apology*, which is devoted entirely to the same purpose. The portrait of Socrates that Xenophon gives in Books III and IV of *Memorabilia* seems, in certain passages, to be heavily influenced by his reading of some of Plato's dialogues, and so the evidentiary value of at least this portion of the work is diminished. Xenophon's *Symposium* is a depiction of

Memorabilia

Socrates in conversation with his friends at a drinking party (it is perhaps inspired by a work of Plato of the same name and character) and is regarded by some scholars as a valuable re-creation of Socrates' thought and way of life. Xenophon's *Oeconomicus* (literally: "estate manager"), a Socratic conversation concerning household organization and the skills needed by the independent farmer, is Xenophon's attempt to bring the qualities he admired in Socrates to bear upon the subject of overseeing one's property. It is unlikely to have been intended as a report of one of Socrates' conversations.

PLATO

Plato, unlike Xenophon, is generally regarded as a philosopher of the highest order of originality and depth. According to some scholars, his philosophical skills made him far better able than Xenophon was to understand Socrates and therefore more valuable a source of information about him. The contrary view is that Plato's originality and vision as a philosopher led him to use his Socratic discourses not as mere devices for reproducing the conversations he had heard but as vehicles for the advocacy of his own ideas (however much they may have been inspired by Socrates) and that he is therefore far more untrustworthy than Xenophon as a source of information about the historical Socrates. Whichever of these two views is correct, it is undeniable that Plato is not only the deeper philosopher but also the greater literary artist. Some of his dialogues are so natural and lifelike in their depiction of conversational interplay that readers must constantly remind themselves that Plato is shaping his material, as any author must.

The role of the character Socrates

Although Socrates is the interlocutor who guides the conversation in most of Plato's dialogues, there are several in which he plays a minor role (*Parmenides*, *Sophist*, *Statesman*, and *Timaeus*, all of which are generally agreed to be among Plato's later works) and one (*Laws*, also composed late) in which he is entirely absent. Why did Plato assign Socrates a small role in some dialogues (and none in *Laws*) and a large role in others? A simple answer is that, by this device, Plato intended to signal to his readers that the dialogues in which Socrates is the major interlocutor convey the philosophy of Socrates, whereas those in which he is a minor figure or does not appear at all present Plato's own ideas.

But there are formidable objections to this hypothesis, and for several reasons most scholars do not regard it as a serious possibility. To begin with, it is unlikely that in so many of his works Plato would have assigned himself so passive and mechanical a role as merely a recording device for the philosophy of Socrates. Furthermore, the portrait of Socrates that results from this hypothesis is not coherent. In some of the dialogues in which he is the principal interlocutor, for example, Socrates insists that he does not have satisfactory answers to the questions he poses—questions such as "What is courage?" (raised in *Laches*), "What is self-control?" (*Charmides*), and "What is piety?" (*Euthyphro*). In other dialogues in which he plays a major role, however, Socrates does offer systematic answers to such questions. In Books II–X of *Republic*, for example, he proposes an elaborate answer to the question, "What is justice?," and in doing so he also defends his view of the ideal society, the condition of the human soul, the nature of reality, and the power of art, among many other topics. Were we to hold that all the Platonic dialogues in which Socrates is the main speaker are depictions of the philosophy of Socrates—a philosophy that Plato endorses but to which he has made no contributions of his own—then we would be committed to the absurd view that Socrates both has and lacks answers to these questions.

For these reasons, there is a broad consensus among scholars that we should not look to works such as *Republic*, *Phaedo*, *Phaedrus*, and *Philebus* for a historically accurate account of the thought of Socrates—even though they contain a speaker called Socrates who argues for certain philosophical positions and opposes others. At the same time, we can explain why Plato uses the literary character of Socrates in many of his writings to present ideas that go well beyond anything that the historical Socrates said or believed. In these works, Plato is developing ideas that were inspired by his encounter with Socrates, using methods of inquiry borrowed from Socrates, and showing how much can be accomplished with these Socratic starting points. That is why he assigns Socrates the role of principal interlocutor, despite the fact that he did not intend these works to be mere re-creations of Socrates' conversations.

Accordingly, the dialogues of Plato that adhere most closely to what he heard from Socrates are those in which the interlocutor called Socrates searches, without apparent success, for answers to questions about the nature of the ethical virtues and other practical topics—works such as *Laches*, *Euthyphro*, and *Charmides*. This does not mean that in these dialogues Plato is not shaping his material or that he is merely writing down, word-for-word, conversations he heard. We cannot know, and it is implausible to suppose, that in these dialogues of unsuccessful search there is a pure rendering of what the historical Socrates said, with no admixture of Platonic interpretation or supplement. All we can reasonably suppose is that here, if anywhere, Plato is re-creating the give-and-take of Socratic conversation, conveying a sense of the methods Socrates used and the assumptions that guided him when he challenged others to defend their ethical ideas and their way of life.

The portrait of Socrates in these dialogues is fully consonant with the one in Plato's *Apology*, and it serves as a valuable supplement to that work. For in the *Apology*, Socrates insists that he does not inquire into natural phenomena ("things in the sky and below the earth"), as Aristophanes alleges. On the contrary, he says, he devotes his life to one question only: how he and others can become good human beings, or as good as possible. The questions he asks others, and discovers that they cannot answer, are posed in the hope that he might acquire greater wisdom about this one subject. This is the Socrates we find in *Laches*, *Euthyphro*, and *Charmides*—but not in *Phaedo*, *Phaedrus*, *Philebus*, or *Republic*. (Or, rather, it is not the Socrates of Books II–X of *Republic*; the portrait of Socrates in Book I is similar in many ways to that in *Apology*, *Laches*, *Euthyphro*, and *Charmides*.) We can therefore say this much about the historical Socrates as he is portrayed in Plato's *Apology* and in some of Plato's dialogues: he has a methodology, a pattern of inquiry, and an orientation toward ethical questions. He can see how misguided his interlocutors are because he is extremely adept at discovering contradictions in their beliefs.

"Socratic method"

"Socratic method" has now come into general usage as a name for any educational strategy that involves cross-examination of students by their teacher. However, the method used by Socrates in the conversations re-created by Plato follows a more specific pattern: Socrates describes himself not as a teacher but as an ignorant inquirer, and the series of questions he asks are designed to show that the principal question he raises (for example, "What is piety?") is one to which his interlocutor has no adequate answer. Typically, the interlocutor is led, by a series of supplementary questions, to see that he must withdraw the answer he at first gave to Socrates' principal question, because that answer falls afoul of the other answers he has given. The method employed by Socrates, in other words, is a strategy for showing that the interlocutor's several answers do not fit together as a group, thus revealing to the interlocutor his own poor grasp of the concepts under discussion. (Euthyphro, for example, in the dialogue named after him, having been asked what piety is, replies that it is whatever is "dear to the gods." Socrates continues to probe, and the ensuing give-and-take can be summarized as follows: Socrates: Are piety and impiety opposites? Euthyphro: Yes. Socrates: Are the gods in disagreement with each other about what is good, what is just, and so on? Euthyphro: Yes. Socrates: So the very same actions are loved by some gods and hated by others? Euthyphro: Yes. Socrates: So those same actions are both pious and impious? Euthyphro: Yes.) The interlocutor, having been refuted by means of premises he himself has agreed to, is free to propose a new answer to Socrates' principal question; or another conversational partner, who has been listening to the preceding dialogue, is allowed to take his place. But al-

though the new answers proposed to Socrates' principal question avoid the errors revealed in the preceding cross-examination, fresh difficulties are uncovered, and in the end the "ignorance" of Socrates is revealed as a kind of wisdom, whereas the interlocutors are implicitly criticized for failing to recognize their ignorance.

Ethical convictions

It would be a mistake, however, to suppose that, because Socrates professes ignorance about certain questions, he suspends judgment about all matters whatsoever. On the contrary, he has some ethical convictions about which he is completely confident. As he tells his judges in his defense speech: human wisdom begins with the recognition of one's own ignorance; the unexamined life is not worth living; ethical virtue is the only thing that matters; and a good human being cannot be harmed (because whatever misfortune he may suffer, including poverty, physical injury, and even death, his virtue will remain intact). But Socrates is painfully aware that his insights into these matters leave many of the most important ethical questions unanswered. It is left to his student Plato, using the Socratic method as a starting point and ranging over subjects that Socrates neglected, to offer positive answers to these questions.

ARISTOTLE

Another important source of information about the historical Socrates—Aristotle—provides further evidence for this way of distinguishing between the philosophies of Socrates and Plato. In 367, some 30 years after the death of Socrates, Aristotle (who was then 17 years old) moved to Athens in order to study at Plato's school, called the Academy. It is difficult to believe that, during his 20 years as a member of that society, Aristotle had no conversations about Socrates with Plato and others who had been personally acquainted with him. There is good reason, then, to suppose that the historical information offered about Socrates in Aristotle's philosophical writings are based on those conversations. What Aristotle tells his readers is that Socrates asked questions but gave no replies, because he lacked knowledge; that he sought definitions of the virtues; and that he was occupied with ethical matters and not with questions about the natural world. This is the portrait of Socrates that Plato's writings, judiciously used, give us. The fact that it is confirmed by Aristotle is all the more reason to accept it.

Life and personality

Although the sources provide only a small amount of information about the life and personality of Socrates, a unique and vivid picture of him shines through, particularly in some of the works of Plato. He was born in or about 470 BC, in Athens. We know the names of his father, Sophroniscus (probably a stonemason), his mother, Phaenarete, and his wife, Xanthippe, and we know that he had three sons. (In Plato's *Theaetetus*, Socrates likens his way of philosophizing to the occupation of his mother, who was a midwife: not pregnant with ideas himself, he assists others with the delivery of their ideas, though they are often stillborn.) With a snub nose and bulging eyes, which made him always appear to be staring, he was unattractive by conventional standards. He served as a hoplite (a heavily armed soldier) in the Athenian army and fought bravely in several important battles. Unlike many of the thinkers of his time, he did not travel to other cities in order to pursue his intellectual interests. Although he did not seek high office, did not regularly attend meetings of the Athenian Assembly (Ecclesia), the city's principal governing body (as was his privilege as an adult male citizen), and was not active in any political faction, he discharged his duties as a citizen, which included not only military service but occasional membership in the Council of Five Hundred, which prepared the Assembly's agenda.

Socrates was not well-born or wealthy, but many of his admirers were, and they included several of the most politically prominent Athenian citizens. When the democratic constitution of Athens was overthrown for a brief time in 403, four years before his trial, he did not leave the city, as did many devoted supporters of democratic rule, including his friend Chaerephon, who had gone to Delphi many years earlier to ask the oracle whether anyone was wiser than Socrates. (The answer was no.)

Socrates, Roman fresco, 1st century BC; at the Ephesus Museum, Seljuk, Turkey.

The expression of same-sex love was not unusual in Athens at this time, and Socrates was physically attracted to beautiful young men. This aspect of his personality is most vividly conveyed in the opening pages of *Charmides* and in the speech of the young and ambitious general Alcibiades at the end of *Symposium*. Socrates' long fits of abstraction, his courage in battle, his resistance to hunger and cold, his ability to consume wine without apparent inebriation, and his extraordinary self-control in the presence of sensual attractions are all described with consummate artistry in the opening and closing pages of *Symposium*.

Socrates' personality was in some ways closely connected to his philosophical outlook. He was remarkable for the absolute command he maintained over his emotions and his apparent indifference to physical hardships. Corresponding to these personal qualities was his commitment to the doctrine that reason, properly cultivated, can and ought to be the all-controlling factor in human life. Thus he has no fear of death, he says in Plato's *Apology*, because he has no knowledge of what comes after it, and he holds that, if anyone does fear death, his fear can be based only on a pretense of knowledge. The assumption underlying this claim is that, once one has given sufficient thought to some matter, one's emotions will follow suit. Fear will be dispelled by intellectual clarity. Similarly, according to Socrates, if one believes, upon reflection, that one should act in a particular way, then, necessarily, one's feelings about the act in question will accommodate themselves to one's belief—one will desire to act in that way. (Thus, Socrates denies the possibility of what has been called "weakness of will"—knowingly acting in a way one believes to be wrong.) It follows that, once one knows what virtue is, it is impossible not to act virtuously. Anyone who fails to act virtuously does so because he incorrectly identifies virtue with something it is not. This is what is meant by the thesis, attributed to Socrates by Aristotle, that virtue is a form of knowledge.

Virtue as a form of knowledge

Socrates' conception of virtue as a form of knowledge explains why he takes it to be of the greatest importance to seek answers to questions such as "What is courage?" and "What is piety?" If we could just discover the answers to these questions, we would have all we need to live our lives well. The fact that Socrates achieved a complete rational

control of his emotions no doubt encouraged him to suppose that his own case was indicative of what human beings at their best can achieve.

But if virtue is a form of knowledge, does that mean that each of the virtues—courage, piety, justice—constitutes a separate branch of knowledge, and should we infer that it is possible to acquire knowledge of one of these branches but not of the others? This is an issue that emerges in several of Plato's dialogues; it is most fully discussed in *Protagoras*. It was a piece of conventional Greek wisdom, and is still widely assumed, that one can have some admirable qualities but lack others. One might, for example, be courageous but unjust. Socrates challenges this assumption; he believes that the many virtues form a kind of unity—though, not being able to define any of the virtues, he is in no position to say whether they are all the same thing or instead constitute some looser kind of unification. But he unequivocally rejects the conventional idea that one can possess one virtue without possessing them all.

The *eirôneia* of Socrates

Another prominent feature of the personality of Socrates, one that often creates problems about how best to interpret him, is (to use the ancient Greek term) his *eirôneia*. Although this is the term from which the English word *irony* is derived, there is a difference between the two. To speak ironically is to use words to mean the opposite of what they normally convey, but it is not necessarily to aim at deception, for the speaker may expect and even want the audience to recognize this reversal. In contrast, for the ancient Greeks *eirôneia* meant "dissembling"—a user of *eirôneia* is trying to hide something. This is the accusation that is made against Socrates several times in Plato's works (though never in Xenophon's). Socrates says in Plato's *Apology*, for example, that the jurors hearing his case will not accept the reason he offers for being unable to stop his philosophizing in the marketplace—that to do so would be to disobey the god who presides at Delphi. (Socrates' audience understood him to be referring to Apollo, though he does not himself use this name. Throughout his speech, he affirms his obedience to the god or to the gods but not specifically to one or more of the familiar gods or goddesses of the Greek pantheon). The cause of their incredulity, he adds, will be their assumption that he is engaging in *eirôneia*. In effect, Socrates is admitting that he has acquired a reputation for insincerity—for giving people to understand that his words mean what they are ordinarily taken to mean when in fact they do not. Similarly, in Book I of *Republic*, Socrates is accused by a hostile interlocutor, Thrasymachus, of "habitual *eirôneia*." Although Socrates says that he does not have a good answer to the question "What is justice?," Thrasymachus thinks that this is just a pose. Socrates, he alleges, is concealing his favoured answer. And in *Symposium*, Alcibiades accuses Socrates of "spending his whole life engaged in *eirôneia* and playing with people" and compares him to a carved figurine whose outer shell conceals its inner contents. The heart of Alcibiades' accusation is that Socrates pretends to care about people and to offer them advantages but withholds what he knows because he is full of disdain.

Plato's portrayal of Socrates as an "ironist" shows how conversation with him could easily lead to a frustrating impasse and how the possibility of resentment was ever present. Socrates was in this sense a masked interlocutor—an aspect of his self-presentation that made him more fascinating and alluring to his audiences but that also added to their distrust and suspicion. And readers, who come to know Socrates through the intervention of Plato, are in somewhat the same situation. Our efforts to interpret him are sometimes not as sound as we would like, because we must rely on judgments, often difficult to justify, about when he means what he says and when he does not.

Even when Socrates goes to court to defend himself against the most serious of charges, he seems to be engaged in *eirôneia*. After listening to the speeches given by his accusers, he says, in the opening sentence of Plato's *Apology*: "I was almost carried away in spite of myself, so persuasively did they speak." Is this the habitual *eirôneia* of Socrates? Or did the speeches of his accusers really have this effect on him? It is difficult to be sure. But, by Socrates' own admission, the suspicion that anything he says might be a pose undermines his ability to persuade the jurors of his good intentions. His *eirôneia* may even have lent support to one of the accusations made against him, that he corrupted the young. For if Socrates really did engage in *eirôneia*, and if his youthful followers delighted in and imitated this aspect of his character, then to that extent he encouraged them to become dissembling and untrustworthy, just like himself.

Background of the trial

The trial of Socrates in 399 BC occurred soon after Athens's defeat at the hands of Sparta in the Peloponnesian War (431–404 BC). Not only were Sparta and Athens military rivals during those years, they also had radically different forms of government. Athens was a democracy: all its adult male citizens were members of the Assembly; many of the city's offices were filled by lot (election was regarded as undemocratic, because it effectively pronounced some citizens better qualified than others); and its citizens enjoyed a high degree of freedom to live and speak as they liked, provided that they obeyed the law and did nothing to undermine the democracy and the public good. Sparta, by contrast, was a mixed regime based on a complex power-sharing arrangement between various elite groups and ordinary citizens, and it exerted far more control than Athens did over education and the daily life of its citizens.

Athenian admirers of Sparta

There was in Athens, particularly among the well-born, wealthy, and young, a degree of admiration for certain aspects of Spartan life and government. These young men, who spent much of their time in the public gymnasia, prided themselves on their toughness, practiced a certain simplicity of style, and grew their hair long—all in imitation of Spartan ways. (As Plato and Xenophon confirm, Socrates himself shared some of these qualities. In Aristophanes' *Birds* [414], the young who express their admiration for Sparta are said to be "Socratizing.") No doubt the fact that Athens, an empire-building city with vast resources and a large population, could not defeat smaller and poorer Sparta—and, in the end, lost its empire to that rival regime—added to the allure of the Spartan political system and way of life.

Ordinary Athenians—people who had to work for a living and did not belong to any of the aristocratic families—were proud of their democratic institutions and the freedoms they enjoyed, and they were well aware that their form of government had internal as well as external enemies and critics. Furthermore, they did not think of civic and religious matters as separate spheres but assumed instead that participation in the religious life of the city, as regulated by democratic institutions, was one of the duties of all citizens and that great harm could come to the city if the gods it recognized were offended or customary religious prohibitions were violated.

RELIGIOUS SCANDAL AND THE COUP OF THE OLIGARCHS

During and soon after the war with Sparta, several events revealed how much damage could be done to Athenian democracy by individuals who did not respect the religious customs of the community, who had no allegiance to the institutions of democracy, or who admired their city's adversary. One night in 415, shortly before a major naval expedition to Sicily was to set sail, many statues of the god Hermes (who protected travelers) were mutilated, presumably by those who wished to prevent the expedition from proceeding. While the matter was being investigated, several men, including one of Socrates' greatest admirers, Alcibiades—who had sponsored and helped to lead the Sicilian expedition—were accused of mocking a religious ceremony and revealing its sacred secrets to outsiders. Some of them were tried and executed. Alcibiades, who had been charged with involvement in other religious scandals before, was called back from Sicily to face trial. The power of his enemies and the suspicion of him was so great, however, that he decided to escape to Sparta rather than return to Athens to face the likelihood of a death sentence. Athens condemned him and his associates to death in absentia, and he proceeded to offer counsel and leader-

ship to Sparta in its fight against Athens. In 407 he returned to Athens and was cleared of the charges against him, though he never fully regained the trust either of the democrats or their opponents. Alcibiades was only one of many followers of Socrates mentioned in Plato's dialogues who were involved in the religious scandals of 415.

In 411 a group of 400 opponents of Athenian democracy staged a coup and tried to install an oligarchy, but they were overthrown in the same year and democracy was restored. Some of them, who were associates of Socrates, went into exile after their revolution failed. In 404, soon after the Athenians' defeat, Sparta installed a group of 30 men (many years later dubbed the Thirty Tyrants) in Athens to establish a far less democratic regime there. The leader of the most extreme wing of this group, Critias, was part of the Socratic circle; so, too, was Charmides, another of the 30. The democrats, many of whom had left Athens when the 30 came to power, defeated them in battle, and democracy was restored the following year. (In Plato's *Apology*, Socrates refers to the reign of the 30 and their unsuccessful attempt to implicate him in their crimes.)

The Thirty Tyrants

THE PERCEIVED FRAGILITY OF ATHENIAN DEMOCRACY

The year in which Socrates was prosecuted, 399, was one in which several other prominent figures were brought to trial in Athens on the charge of impiety. That is unlikely to have been a coincidence; rather, it suggests that there was, at the time, a sense of anxiety about the dangers of religious unorthodoxy and about the political consequences that religious deviation could bring. Two attempts to put an end to Athenian democracy had occurred in recent years, and the religious scandals of 415 were not so far in the past that they would have been forgotten. Because a general amnesty had been negotiated, no one, except the 30 and a few others, could be tried for offenses committed prior to 403, when the 30 were defeated. But this would not have prevented an accusation from being brought against someone who committed a crime after 403. If Socrates had continued, during the years after 403, to engage in the same practices that were so characteristic of him throughout his adult life, then not even the most ardent supporters of the amnesty would have objected to bringing him to trial. And once a trial had begun, it was common practice for prosecutors to mention anything that might be judged prejudicial to the accused. There was no legal custom or court-appointed judge that would have prevented Socrates' accusers from referring to those of his admirers—Alcibiades, Critias, Charmides, and the like—who at one time had been enemies of democratic Athens or had been associated with religious scandal. The law that Socrates was alleged to have violated was a law against impiety, but in support of that accusation he also was accused of having corrupted the young. His jury might have taken his association with opponents of the democracy, or with persons convicted or suspected of religious crimes, to be grounds for considering him a dangerous man.

The fact that one of those who assisted in the prosecution of Socrates and spoke against him—Anytus—was a prominent democratic leader makes it all the more likely that worries about the future of Athenian democracy lay behind Socrates' trial. And even if neither Anytus nor the other prosecutors (Meletus and Lycon) harboured such fears, it is hard to believe that they were entirely absent from the minds of those who heard his case. In any event, because Socrates openly displayed his antidemocratic ideas in his defense speech, it would have been difficult for jurors to set aside his association with opponents of the democracy, even if they had been inclined to do so. Athenian democracy must have seemed extremely fragile in 399. It is only with the benefit of hindsight that we can see that its institutions were strong enough to last most of the rest of the 4th century.

It is not known with certainty whether those who prosecuted Socrates mentioned Alcibiades and Critias at his trial—there is no record of their speeches, and it is difficult to interpret the evidence about what they did say. But it is very likely that specific names were mentioned. In Plato's *Apology*, Socrates notes that his accusers alleged of certain individuals that they were his students, an accusation he lamely denies on the grounds that, because he has never undertaken to teach anyone, he cannot have had students. Furthermore, Xenophon reports in *Memorabilia* that, according to "the accuser," Alcibiades and Critias were followers of Socrates. The word *accuser* is taken by some scholars to be a reference to one of the three persons who spoke against Socrates in 399, though others take Xenophon to be defending Socrates against charges made against him in a pamphlet written several years later by Polycrates, a teacher of rhetoric. In any event, many years later, in the 4th century, the orator Aeschines (390–314 BC), in his speech "Against Timarchus," asserted in public that Socrates was convicted because he was "shown to have been the teacher of Critias, one of the thirty who had overthrown the democracy."

But even if Socrates' association with Critias and Alcibiades was an important factor leading to his trial and conviction, it certainly was not the only ingredient of the case against him, nor even the most important one. The law that Socrates was alleged to have violated was a law against impiety, and the thrust of his defense, as presented by Plato, was that his life has been consumed by his single-minded devotion to the god. The Socrates who speaks to us in Plato's *Apology* has no doubt that the charge of impiety against him must be refuted. There is no reason to suspect that this charge was a mere pretext and that what Socrates was really being prosecuted for was his antidemocratic associations and ideas. The political background of his trial is important because it helps to explain why he was not prosecuted in the 430s or 420s or at any other time of his life. Everything known about him indicates that he was the same man, and lived the same sort of life, in 399 and in 423, the year of *Clouds*. What made him the object of prosecution in 399, after so many years during which his behaviour was tolerated, was a change in political circumstances. But it remains the case, according to the Socrates of *Apology*, that his alleged religious unorthodoxy was deeply worrying to his prosecutors and jurors. That is why this allegation receives all his attention.

THE ATHENIAN IDEAL OF FREE SPEECH

That Socrates was prosecuted because of his religious ideas and political associations indicates how easily an ideal held dear by his fellow Athenians—the ideal of open and frank speech among citizens—could be set aside when they felt insecure. This ideal and its importance in Athens are well illustrated by the remark of the orator Demosthenes, that in Athens one is free to praise the Spartan constitution, whereas in Sparta it is only the Spartan constitution that one is allowed to praise. Were there other instances, besides the trial of Socrates, in which an Athenian was prosecuted in court because of the dangerous ideas he was alleged to have circulated? Centuries after Socrates' death, several writers alleged that many other intellectual figures of his time—including Protagoras, Anaxagoras, Damon, Aspasia, and Diagoras—were exiled or prosecuted. Several scholars have concluded that Athens's allegiance to the ideal of freedom of speech was deeply compromised during the last decades of the 5th century. Others have argued that much or all of the evidence for a period of persecution and harassment was invented by writers who wanted to claim, as a badge of honour for their favourite philosophers, that they, too, like the universally admired Socrates, had been persecuted by the Athenians. What can safely be said is this: the trial of Socrates is the only case in which we can be certain that an Athenian was legally prosecuted not for an overt act that directly harmed the public or some individual—such as treason, corruption, or slander—but for alleged harm indirectly caused by the expression and teaching of ideas.

According to Plato's *Apology*, the vote to convict Socrates was very close: had 30 of those who voted for conviction cast their ballots differently, he would have been acquitted. (So he was convicted by a majority of 59. Assuming, as many scholars do, that the size of his jury was 501, 280 favoured conviction and 221 opposed it.) It is reasonable to speculate that many of those who opposed conviction did so partly because, however little they cared for what

The vote to convict Socrates

Socrates thought and how he lived, they cherished the freedom of speech enjoyed by all Athenians and attached more importance to this aspect of their political system than to any harm Socrates may have done in the past or might do in the future. The Athenian love of free speech allowed Socrates to cajole and criticize his fellow citizens for the whole of his long life but gave way—though just barely—when it was put under great pressure.

Plato's *Apology*

Although in none of Plato's dialogues is Plato himself a conversational partner or even a witness to a conversation, in the *Apology* Socrates says that Plato is one of several friends in the audience. In this way Plato lets us know that he was an eyewitness of the trial and therefore in the best possible position to write about it. The other account we have of the trial, that of Xenophon, a contemporary of Socrates, is of a very different character. We know that Xenophon was not present as a live witness. He tells his readers that he is reporting only a portion of Socrates' speech and that he learned about the trial from Hermogenes, a member of the Socratic circle.

It is not surprising, then, that there are significant differences between Plato's and Xenophon's accounts of what was said at the trial. (Xenophon, for example, dwells on the troubles of old age from which Socrates is escaping by being condemned to death, whereas Plato barely alludes to Socrates' age.) Of greater importance is the fact that the two *Apologys* agree in many details. They agree about what the charges against Socrates were: failing to acknowledge the gods recognized by the city, introducing other new divinities, and corrupting the young. They also agree that Meletus supported his accusation by referring to a divine voice or sign that Socrates claimed as his personal guide; that Socrates acknowledged the guidance of this divine sign in his speech; that part of Socrates' defense consisted of a cross-examination of Meletus; that Socrates referred to an inquiry made by his friend, Chaerephon, to the Delphic oracle; that the response of the oracle confirmed that a unique status had been conferred upon Socrates by the god; that, having been found guilty, Socrates refused to propose a punishment that the jury would find acceptable; and that, after the jury voted in favour of the death penalty, he once again addressed the jury and expressed no regrets for his manner of living or the course of his trial. There is no reason to suppose that Xenophon had learned of these aspects of the trial from Plato. His agreement with Plato about these matters assures us that they are not fabrications.

But can we go so far as to say that in Plato's *Apology* there is a word-for-word transcription (or something close to it) of the speech Socrates gave in his defense? It would not have been impossible for Plato to have managed such a feat by taking extensive notes, comparing his memory with that of others, and gradually perfecting a rendition that aimed at replicating the original as closely as possible. Unfortunately, there is no way to prove that Plato was striving to achieve this kind and degree of accuracy. Some scholars, in fact, have argued that Plato was engaged in a much different project: his *Apology*, they have noted, is similar in many respects to the works of contemporary orators and teachers of rhetoric—in particular to a rhetorical exercise, "Defense of Palamades," by Gorgias—and they infer that in composing the *Apology* in this fashion Plato was not seeking historical accuracy but instead striving to outdo or to parody the orators for whom he felt disdain. But this hypothesis is just as speculative as the supposition that Plato strove to record as accurately as possible the actual speech of Socrates.

We cannot eliminate the possibility that some parts of the speech Plato wrote were not actually delivered at the trial or were expressed rather differently. Plato's speech represents his creative attempt to defend Socrates and his way of life and to condemn those who voted to kill him. In fact, Plato's motives in writing the *Apology* are likely to have been complex. One of them, no doubt, was to defend and praise Socrates by making use of many of the points Socrates himself had offered in his speech. But, as any reader of the work can see, Plato is at the same time using the trial and death of Socrates to condemn Athens, to call upon his readers to reject the conventional life that Athens would have preferred Socrates to lead, and to choose instead the life of a Socratic philosopher. In the 4th century BC Athens had no norm of accurate reportage or faithful biography, and so Plato would have felt free to shape his material in whatever way suited his multiple aims. Because it was Socrates he wished to praise, he had no choice but to make the Socrates of the *Apology* close to the original. But he would not have felt bound merely to reproduce, as best he could, the speech that Socrates delivered.

Plato's motives in the *Apology*

In any event, the historical accuracy of Plato's *Apology* should not be the only question on the reader's mind. Of equal importance is whether Plato's Socrates really is guilty

"The Death of Socrates," oil on canvas by Jacques-Louis David, 1787; in the Metropolitan Museum of Art, New York.

of the charges brought against him, whether he is a wholly just and admirable person, whether his manner of living is the one that is most worthwhile (or perhaps even the only one that is worthwhile at all, as Socrates insists), and whether there is any reason for a political community to be concerned about the harm such a person might do. Surely the last thing Plato would have wanted his readers to do with the *Apology* is to ignore its philosophical, religious, and political dimensions in order to concentrate solely on its accuracy as a piece of historical reportage.

The public's hatred of Socrates

Part of the fascination of Plato's *Apology* consists in the fact that it presents a man who takes extraordinary steps throughout his life to be of the greatest possible value to his community but whose efforts, far from earning him the gratitude and honour he thinks he deserves, lead to his condemnation and death at the hands of the very people he seeks to serve. Socrates is painfully aware that he is a hated figure and that this is what has led to the accusations against him. He has little money and no political savvy or influence, and he has paid little attention to his family and household—all in order to serve the public that now reviles him. What went wrong?

THE IMPRESSION CREATED BY ARISTOPHANES

Socrates goes to some length to answer this question. Much of his defense consists not merely in refuting the charges but in offering a complex explanation of why such false accusations should have been brought against him in the first place. Part of the explanation, he believes, is that he has long been misunderstood by the general public. The public, he says, has focused its distrust of certain types of people upon him. He claims that the false impressions of his "first accusers" (as he calls them) derive from a play of Aristophanes (he is referring to *Clouds*) in which a character called Socrates is seen "swinging about, saying he was walking on air and talking a lot of nonsense about things of which I know nothing at all." The Socrates of Aristophanes' comedy is the head of a school that investigates every sort of empirical phenomenon, regards clouds and air as divine substances, denies the existence of any gods but these, studies language and the art of argument, and uses its knowledge of rhetorical devices to "make the worse into the stronger argument," as the Socrates of the *Apology* puts it in his speech. Socrates' corruption of the young is also a major theme of *Clouds*: it features a father (Strepsiades) who attends Socrates' school with his son (Pheidippides) in order to learn how to avoid paying the debts he has incurred because of his son's extravagance. In the end, Pheidippides learns all too well how to use argumentative skills to his advantage; indeed, he prides himself on his ability to prove that it is right for a son to beat his parents. In the end, Strepsiades denounces Socrates and burns down the building that houses his school.

To "make the worse into the stronger argument"

This play, Socrates says, has created the general impression that he studies celestial and geographic phenomena and, like the Sophists who travel from city to city, takes a fee for teaching the young various skills. Not so, says Socrates. He thinks it would be a fine thing to possess the kinds of knowledge these Sophists claim to teach, but he has never discussed these matters with anyone—as his judges should be able to confirm for themselves, because, he says, many of them have heard his conversations.

THE HUMAN RESISTANCE TO SELF-REFLECTION

But this can only be the beginning of Socrates' explanation, for it leads to further questions. Why should Aristophanes have written in this way about Socrates? The latter must have been a well-known figure in 423, when *Clouds* was produced, for Aristophanes typically wrote about and mocked figures who already were familiar to his audience. Furthermore, if, as Socrates claims, many of his jurors had heard him in discussion and could therefore confirm for themselves that he did not study or teach others about clouds, air, and other such matters and did not take a fee as the Sophists did, then why did they not vote to acquit him of the charges by an overwhelming majority?

Socrates provides answers to these questions. Long before Aristophanes wrote about him, he had acquired a reputation among his fellow citizens because he spent his days attempting to fulfill his divine mission to cross-examine them and to puncture their confident belief that they possessed knowledge of the most important matters. Socrates tells the jurors that, as a result of his inquiries, he has learned a bitter lesson about his fellow citizens: not only do they fail to possess the knowledge they claim to have, but they resent having this fact pointed out to them, and they hate him for his insistence that his reflective way of life and his disavowal of knowledge make him superior to them. The only people who delight in his conversation are the young and wealthy, who have the leisure to spend their days with him. These people imitate him by carrying out their own cross-examinations of their elders. Socrates does admit, then, that he has, to some degree, set one generation against another—and in making this confession, he makes it apparent why some members of the jury may have been convinced, on the basis of their own acquaintance with him, that he has corrupted the city's young.

One of the most subtle components of Socrates' explanation for the hatred he has aroused is his point that people hide the shame they feel when they are unable to withstand his destructive arguments. His reputation as a corrupter of the young and as a Sophist and an atheist is sustained because it provides people with an ostensibly reasonable explanation of their hatred of him. No one will say, "I hate Socrates because I cannot answer his questions, and he makes me look foolish in front of the young." Instead, people hide their shame and the real source of their anger by seizing on the general impression that he is the sort of philosopher who casts doubt on traditional religion and teaches people rhetorical tricks that can be used to make bad arguments look good. These ways of hiding the source of their hatred are all the more potent because they contain at least a grain of truth. Socrates, as both Plato and Xenophon confirm, is a man who loves to argue: in that respect he is like a Sophist. And his conception of piety, as revealed by his devotion to the Delphic oracle, is highly unorthodox: in that respect he is like those who deny the existence of the gods.

Socrates believes that this hatred, whose real source is so painful for people to acknowledge, played a crucial role in leading Meletus, Anytus, and Lycon to come forward in court against him; it also makes it so difficult for many members of the jury to acknowledge that he has the highest motives and has done his city a great service. Aristophanes' mockery of Socrates and the legal indictment against him could not possibly have led to his trial or conviction were it not for something in a large number of his fellow Athenians that wanted to be rid of him. This is a theme to which Socrates returns several times. He compares himself, at one point, to a gadfly who has been assigned by the god to stir a large and sluggish horse. Note what this implies: the bite of the fly cannot be anything but painful, and it is only natural that the horse would like nothing better than to kill it. After the jury has voted in favour of the death penalty, Socrates tells them that their motive has been their desire to avoid giving a defense of their lives. Something in people resists self-examination: they do not want to answer deep questions about themselves, and they hate those who cajole them for not doing so or for doing so poorly. At bottom, Socrates thinks that all but a few people will strike out against those who try to stimulate serious moral reflection in them. That is why he thinks that his trial is not merely the result of unfortuitous events—a mere misunderstanding caused by the work of a popular playwright—but the outcome of psychological forces deep within human nature.

Socrates as a gadfly

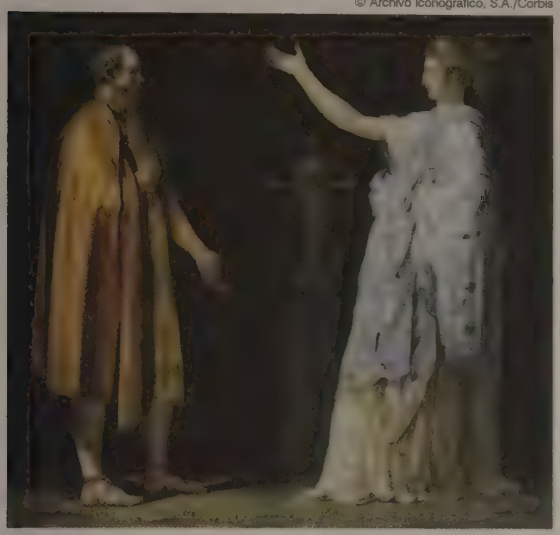
SOCRATES' CRITICISM OF DEMOCRACY

Socrates' analysis of the hatred he has incurred is one part of a larger theme that he dwells on throughout his speech. Athens is a democracy, a city in which the many are the dominant power in politics, and it can therefore be expected to have all the vices of the many. Because most people hate to be tested in argument, they will always take action of some sort against those who provoke them with

questions. But that is not the only accusation Socrates brings forward against his city and its politics. He tells his democratic audience that he was right to have withdrawn from political life, because a good person who fights for justice in a democracy will be killed. In his cross-examination of Meletus, he insists that only a few people can acquire the knowledge necessary for improving the young of any species, and that the many will inevitably do a poor job. He criticizes the Assembly for its illegal actions and the Athenian courts for the ease with which matters of justice are distorted by emotional pleading. Socrates implies that the very nature of democracy makes it a corrupt political system. Bitter experience has taught him that most people rest content with a superficial understanding of the most urgent human questions. When they are given great power, their shallowness inevitably leads to injustice.

The charge of impiety

Socrates spends a large part of his speech trying to persuade his fellow citizens that he is indeed a pious man, because his philosophical mission has been carried out in obedience to the god. It is remarkable that this is nearly the



Detail from "Socrates and Philosophy," tempera painting by Antonio Canova; in the Gipsoteca Canoviana, Possagno, Italy.

only positive argument he offers, in Plato's *Apology*, to support his claim that he is a pious man. The only other evidence he supplies is introduced only because Meletus, upon cross-examination, asserts that Socrates believes that there are no gods or divinities at all, an accusation far more sweeping than—and indeed contradictory to—the official indictment, which asserted that Socrates did not acknowledge the gods recognized by the city but instead believed in different and new gods. Socrates quickly points out the absurdity of this new accusation. Meletus, he notes, has referred in his speech to a certain strange divinity (*daimon*) who comes to Socrates to give him advice. Presumably Meletus has offered this as evidence that Socrates believes in new gods that are different from the ones generally recognized in Athens. But if Meletus admits that Socrates is guided by a divine being, then he cannot be taken seriously when he also says that Socrates is a complete atheist.

SOCRATES' RADICAL RECONCEPTION OF PIETY

Conventional forms of piety

These two modes of Socrates' religiosity—serving the god by cross-examining one's fellow citizens and accepting the guidance of a divine voice—are nothing like the conventional forms of piety with which Socrates' contemporaries were familiar. The Athenians, like all Greeks in the ancient world, expressed their piety by participating in festivals, making sacrifices, visiting shrines, and the like. They assumed that it was the better part of caution to show one's devotion to the gods in these public and conventional ways because, if the gods were not honoured, they could easily harm or destroy even the best of men and women and their families and cities as well. The Socrates of Plato's *Apology* does not refer to his participation in these ceremonies and rituals. (The Socrates of Xenophon's *Apology* does, however, and, in this and many other ways, Plato's Socrates is the more unconventional and provocative of the two and a figure more likely to be hated and feared.) It is impossible to know whether the historical Socrates participated fully (or at all) in conventional forms of religious observance, but, if Plato's account of his philosophy is accurate, then Socrates lacked the typical Athenian's motives for doing so. He cannot believe that the gods might harm him, because he is confident that he is a good man and that a good man cannot be harmed. That is why he has no fear of other human beings. Even if the jury votes to banish him from Athens or to kill him, he will not be worse off, because his peculiar kind of wisdom and virtue—his acknowledgment of his ignorance and commitment to continual self-examination—will remain intact. That is also why he is sure that, when he dies, his affairs will not be neglected by the gods. They must be entirely benign in their attitude toward someone like him, who has served them so well, and so he has no need to offer them gifts, if gifts are a device for incurring their favour or protecting oneself from their destructive power.

In effect, then, Socrates admits that his understanding of piety is radically different from the conventional conception. In keeping with his conception of virtue as a form of knowledge he uses an intellectual test, not merely a ceremonial test, to determine whether someone is pious. You may participate in the conventional practices of civic religion, but can you say what piety is? If you cannot, do you at least admit your ignorance and search constantly for a better understanding of piety, as the god wishes you to do? More generally, though you may think you are a good person, can you say what your virtues consist of? If you cannot, and if you do not spend your life trying, then your goodness is a sham.

Socrates' *daimon*

Socrates' reconception of piety must have struck his fellow citizens as all the more bizarre and threatening because it was accompanied by his unapologetic and grateful acceptance of the divine sign, which Meletus ridicules—a voice that has come to him since childhood, warning him away from certain undertakings and in doing so giving him unfailing advice. In Xenophon's *Apology*, Socrates seeks to portray the *daimon* that guides him as a phenomenon akin to others with which his fellow citizens are quite familiar: "Those who rely on bird-calls and the utterances of men are, I suppose, receiving guidance from voices. Can there be any doubt that thunder has a voice or that it is an omen of the greatest significance?" But an Athenian of conventional piety would have been able to spot the weakness of this attempt to assimilate Socrates' divine voice to the experience of a seer who makes predictions based on the interpretation of natural phenomena. Such seers were appointed and regulated by civic procedures. Socrates was not designated by the city to serve in an official religious capacity, and therefore, in claiming to have experiences that put him directly in touch with the divine, he was circumventing the normal route by which citizens gained access to the sources of religious inspiration. The Socrates of Plato's *Apology*, unlike that of Xenophon's, makes no attempt to portray his divine sign as a phenomenon that can create no rift or distance between himself and others. On the contrary, he attributes his decision not to participate in the political life of the community beyond the minimal duties of citizenship to the influence of his divine sign, and he is confident that his decision to come to court and contest the charges against him (leaving the city and living in exile was an option) was the right one because it was not opposed by the divine sign. The *daimon* Socrates listens to is a divinity that makes a political difference: it tells him what kind of relationship he should have with his fellow citizens and how he should conduct himself in public affairs. Thus, not only does Socrates have an unorthodox conception of piety and of what the gods want from the citizens of the city, but also he claims to receive infallible guidance from a voice that does not hesitate to speak to him about public matters.

THE DANGER POSED BY SOCRATES

An open-minded and conscientious member of the jury could therefore have come to the conclusion that Socrates posed a significant threat to the city and should be found guilty of the charges against him. In a way, Socrates did fail to acknowledge the gods recognized by the city, he did introduce new gods, and, by teaching these things to the young who gathered around him, he did corrupt them. He may have referred to "the god" or "the gods," but his conception of what is involved in attending to the gods was utterly novel and politically dangerous. The fact that Socrates saw his piety as the genuine article, and the unreflective virtue of his fellow citizens as false virtue, indicates that he took the entire religious life of Athens, no less than its political life, to be unworthy of a good man.

If there is any doubt that the unorthodox form of piety Socrates embodies could have brought him into direct conflict with the popular will, one need only think of the portion of Plato's *Apology* in which Socrates tells the jurors that he would obey the god rather than them. Imagining the possibility that he is acquitted on the condition that he cease philosophizing in the marketplace, he unequivocally rejects the terms of this hypothetical offer, precisely because he believes that his religious duty to call his fellow citizens to the examined life cannot be made secondary to any other consideration: "Men of Athens, I salute you and hold you dear, but I will obey the god rather than you, and so long as I take breath and am able, I will never cease philosophizing." But there was no need for him to have admitted, in such explicit terms, that his conception of piety might require him, in certain circumstances, to disobey a civic order. It is characteristic of his entire speech that he brings into the open how contemptuous he is of Athenian civic life and his fellow citizens. He prides himself on the fact that he will say nothing to curry favour with the jurors or to conceal his attitude of superiority to them—even though he realizes that this is likely to lead some of them to vote against him out of resentment. Others may throw themselves on the pity of the jury or bring their tearful children and friends to court; but these typical modes of behaviour corrupt the legal system, and Socrates will not stoop to such tactics. Here, as in so many parts of his speech, he treats his day in court as an opportunity to counter-indict his accusers and his fellow citizens (those, at any rate, who voted against him) for the way they lead their lives. (Another example: after he has been found guilty and has the opportunity to propose a punishment, he tells the jury that he should receive free meals for the remainder of his life, because this is what he deserves—though in the end he offers to pay one mina of silver, equivalent to about one hundred days' wages, a penalty that his wealthy friends attending the trial increase to 30 minas.) In effect, Socrates uses the occasion of his trial to put his accusers and the jurors on trial. But this was a natural role for him, because he had done the same thing, day after day, to everyone he met.

Socrates versus Plato

We can conclude that Plato was not blind to the civic and religious dangers created by Socrates. Part of what makes his *Apology* so complex and gripping is that it is not a one-sided encomium that conceals the features of the Socratic way of life that lay behind the anxiety and resentment felt by many of his fellow citizens. Plato, of course, leaves no doubt that he sides with Socrates and against Athens, but in doing so he allows us to see why Socrates had enemies as well as friends. The multisidedness of Plato's portrait adds to its verisimilitude and should increase our confidence in him as a source of our understanding of the historical Socrates. A defense of Socrates that portrayed him as an innocuous preacher of moral pieties would have left us wondering why he was sentenced to death, and indeed why anyone bothered to indict him in the first place.

The multisidedness of Plato's portrait

Plato gives no hint in his *Apology* that he had any reservations about the way Socrates led his life or the doctrines that guided him; the format of the *Apology* prevents him from doing so. He has made the decision to let Socrates speak for himself in this work and to refrain from offering any of his own reflections on the justice or injustice of the charges against his teacher. But, in the *Republic*, he puts into the mouth of its principal interlocutor, "Socrates," an observation about the corrosive power that philosophy can have when it takes hold at too early an age. When young people first hear philosophical questions about the traditional moral standards they have learned from their parents and their community, and when they see that it is difficult to defend these orthodoxies without falling into contradiction, they are prone to reject all traditional morality and to become essentially lawless. For this reason, philosophy may come to be seen as a dangerous and disreputable pursuit. The Socrates of the *Republic* therefore suggests that in an ideal society the young should not be exposed to ethical doubt until they are well into their maturity. This, of course, is not a restriction that the historical Socrates imposed on himself. In Plato's *Apology*, Socrates prides himself on addressing his questions to every Athenian—no one, in his view, is too young or too old for the examined life—and he freely acknowledges that the young love to see their elders embarrassed when they are unable to defend their beliefs. Whereas the Socrates of Plato's *Apology* assumes that there is no need to place limits on philosophical inquiry, the Socrates of the *Republic*—who speaks as the mouthpiece of Plato—holds that in an ideal society this kind of activity would be carefully regulated. Similarly, in Plato's *Laws*, the main speaker, an unnamed visitor from Athens, praises Sparta and Crete for forbidding the young to criticize the laws of their communities. Plato's great admiration for Socrates was all the more remarkable because it coexisted not only with a recognition of why Socrates was considered dangerous but also with his belief that Socrates was, to some degree, guilty of impiety and of corrupting the young.

The Socrates of the *Republic*

The legacy of Socrates

Socrates' thought was so pregnant with possibilities, his mode of life so provocative, that he inspired a remarkable variety of responses. One of his associates, Aristippus of Cyrene—his followers were called "Cyrenaics," and their school flourished for a century and a half—affirmed that pleasure is the highest good. (Socrates seems to endorse this thesis in Plato's *Protagoras*, but he attacks it in *Gorgias* and other dialogues.) Another prominent follower of Socrates in the early 4th century BC, Antisthenes, emphasized the Socratic doctrine that a good man cannot be harmed; virtue, in other words, is by itself sufficient for happiness. That doctrine played a central role in a school of thought, founded by Diogenes of Sinope, that had an enduring influence on Greek and Roman philosophy: Cynicism. Like Socrates, Diogenes was concerned solely with ethics, practiced his philosophy in the marketplace, and upheld an ideal of indifference to material possessions, political power, and conventional honours. But the Cynics, unlike Socrates, treated all conventional distinctions and cultural traditions as impediments to the life of virtue. They advocated a life in accordance with nature and regarded animals and human beings who did not live in societies as being closer to nature than contemporary human beings. (The term *cynic* is derived from the Greek word for dog. Cynics, therefore, live like beasts.) Starting from the Socratic premise that virtue is sufficient for happiness, they launched attacks on marriage, the family, national distinctions, authority, and cultural achievements. But the two most important ancient schools of thought that were influenced by Socrates were Stoicism, founded by Zeno of Citium, and Skepticism, which became, for many centuries, the reigning philosophical stance of Plato's Academy after Arcesilaus became its leader in 273 BC. The influence of Socrates on Zeno was mediated by the Cynics, but Roman Stoics—particularly Epictetus—regarded Socrates as the paradigm of sagacious inner strength, and they invented new arguments for the Socratic thesis that virtue is sufficient for happiness. The Stoic doctrine that divine intelligence pervades the world and rules for the best borrows heavily from ideas attributed to Socrates by Xenophon in the *Memorabilia*.

Stoicism and Skepticism

Like Socrates, Arcesilaus wrote nothing. He philosophized by inviting others to state a thesis; he would then prove, by Socratic questioning, that their thesis led to a contradiction. His use of the Socratic method allowed Arcesilaus and his successors in the Academy to hold that they were remaining true to the central theme of Plato's writings. But, just as Cynicism took Socratic themes in a direction Socrates himself had not developed and indeed would have rejected, so, too, Arcesilaus and his Skeptical followers in Plato's Academy used the Socratic method to advocate a general suspension of all convictions whatsoever and not merely a disavowal of knowledge. The underlying thought of the Academy during its Skeptical phase is that because there is no way to distinguish truth from falsity, we must refrain from believing anything at all. Socrates, by contrast, merely claims to have no knowledge, and he regards certain theses as far more worthy of our credence than their denials.

Although Socrates exerted a profound influence on Greek and Roman thought, not every major philosopher of antiquity regarded him as a moral exemplar or a major thinker. Aristotle approves of the Socratic search for definitions but criticizes Socrates for an overintellectualized conception of the human psyche. The followers of Epicurus, who were philosophical rivals of the Stoics and Academics, were contemptuous of him.

Giraudon/Art Resource, New York

Socrates and his students, miniature from al-Mubashshir, *Mukhtar al-Hikam* ("The Better Sentences and Most Precise Dictions"), Seljuk manuscript, early 13th century; in the Topkapı Palace Museum, Istanbul.

With the ascendancy of Christianity in the medieval period, the influence of Socrates was at its nadir: he was, for many centuries, little more than an Athenian who had been condemned to death. But when Greek texts, and thus the works of Plato, the Stoics, and the Skeptics, became increasingly available in the Renaissance, the thought and personality of Socrates began to play an important role in European philosophy. From the 16th to the 19th century the instability and excesses of Athenian democracy became a common motif of political writers; the hostility of Xenophon and Plato, fed by the death of Socrates, played an important role here. Comparisons between Socrates and Christ became commonplace, and they remained so even into the 20th century—though the contrasts drawn between them, and the uses to which their similarities were put, varied greatly from one author and period to another. The divine sign of Socrates became a matter of controversy: was he truly inspired by the voice of God? Or was the sign only an intuitive and natural grasp of virtue? (So thought Montaigne.) Did he intend to undermine the irrational and merely conventional aspects of religious practice and thus to place religion on a scientific footing? (So thought the 18th-century Deists.)

Comparisons between Socrates and Christ

In the 19th century Socrates was regarded as a seminal figure in the evolution of European thought or as a Christ-like herald of a higher existence. G.W.F. Hegel saw in Socrates a decisive turn from pre-reflective moral habits to a self-consciousness that, tragically, had not yet learned how to reconcile itself to universal civic standards. Søren Kierkegaard, whose dissertation examined Socratic irony, found in Socrates a pagan anticipation of his belief that Christianity is a lived doctrine of almost impossible demands; but he also regarded Socratic irony as a deeply flawed indifference to morality. Friedrich Nietzsche struggled throughout his writings against the one-sided rationalism and the destruction of cultural forms that he found in Socrates.

In contrast, in Victorian England Socrates was idealized by utilitarian thinkers as a Christ-like martyr who laid the foundations of a modern, rational, scientific worldview. John Stuart Mill mentions the legal executions of Socrates and of Christ in the same breath in order to call attention to the terrible consequences of allowing common opinion to persecute unorthodox thinkers. Benjamin Jowett, the principal translator of Plato in the late 19th century, told his students at Oxford, "The two biographies about which we are most deeply interested (though not to the same degree) are those of Christ and Socrates." Such comparisons continued into the 20th century: Socrates is treated as a "paradigmatic individual" (along with Buddha, Confucius, and Christ) by the German existentialist philosopher Karl Jaspers.

Influence in the 20th century

The conflict between Socrates and Athenian democracy shaped the thought of 20th-century political philosophers such as Leo Strauss, Hannah Arendt, and Karl Popper. The tradition of self-reflection and care of the self initiated by Socrates fascinated Michel Foucault in his later writings. Analytic philosophy, an intellectual tradition that traces its origins to the work of Gottlob Frege, G.E. Moore, and Bertrand Russell in the late 19th and early 20th century, uses, as one of its fundamental tools, a process called "conceptual analysis," a form of nonempirical inquiry that bears some resemblance to Socrates' search for definitions.

But the influence of Socrates is felt not only among philosophers and others inside the academy. He remains, for all of us, a challenge to complacency and a model of integrity. (Ri.Kr.)

BIBLIOGRAPHY

General studies. Overviews are presented in C.C.W. TAYLOR, *Socrates* (1998; also reissued as C.C.W. TAYLOR, R.M. HARE, and JONATHAN BARNES, *Greek Philosophers: Socrates, Plato, and Aristotle*, 1999); THOMAS C. BRICKHOUSE and NICHOLAS D. SMITH, *The Philosophy of Socrates* (2000); and W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 3 (1969; also reissued as *The Sophists*, 1971), pp. 323–507. A large scholarly literature focuses on the seminal work of GREGORY VLASTOS, including his *Socrates: Ironist and Moral Philosopher* (1991), and his *Socratic Studies* (1994). CHARLES H. KAHN, *Plato and the Socratic Dialogue: The Philosophical Use of a Literary Form* (1996, reissued 1998), argues that Plato's dialogues are devoid of evidence of a distinctive Socratic philosophy. Discussions of many diverse aspects of the Socrates of Plato's early dialogues are included in THOMAS C. BRICKHOUSE and NICHOLAS D. SMITH, *Plato's Socrates* (1994); TERENCE IRWIN, *Plato's Ethics*, chapters 1–9 (1995), pp. 3–147; GERASIMOS XENOPHON SANTAS, *Socrates: Philosophy in Plato's Early Dialogues* (1979, reissued in 1982; also published as *Socrates*, 1999); and KENNETH SEESKIN, *Dialogue and Discovery: A Study in Socratic Method* (1987). Socratic irony is discussed in ALEXANDER NEHAMAS, *The Art of Living: Socratic Reflections from Plato to Foucault* (1998), pp. 19–98. PAUL A. VANDER WAERDT (ed.), *The Socratic Movement* (1994), contains many essays on the non-Platonic "Socratic discourses" and the philosophical movements inspired by Socrates in antiquity. Further discussion of the problem of discovering the historical Socrates may be found in DEBRA NAILS, *Agora, Academy, and the Conduct of Philosophy* (1995). Several chapters of R.B. RUTHERFORD, *The Art of Plato: Ten Essays in Platonic Interpretation* (1995), discuss our knowledge of Socrates and literary aspects of Plato's early dialogues. An unusual perspective is presented in JOHN BEVERSLUIS, *Cross-Examining Socrates: A*

Defense of the Interlocutors in Plato's Early Dialogues (2000). HUGH H. BENSON, *Socratic Wisdom: The Model of Knowledge in Plato's Early Dialogues* (2000), examines the Socratic method and epistemology.

Among several anthologies of articles about Socrates are WILLIAM J. PRIOR (ed.), *Socrates: Critical Assessments*, 4 vol. (1996); HUGH H. BENSON (ed.), *Essays on the Philosophy of Socrates* (1992); MICHAEL C. STOKES and BARRY S. GOWER (eds.), *Socratic Questions: New Essays on the Philosophy of Socrates and Its Significance* (1992); and GREGORY VLASTOS (ed.), *The Philosophy of Socrates* (1971, reprinted 1980).

Other noteworthy studies are J.K. ANDERSON, *Xenophon* (1974, reissued 2001); ANTON-HERMANN CHROUST, *Socrates, Man and Myth: The Two Socratic Apologies of Xenophon* (1957); LEO STRAUSS, *Xenophon's Socrates* (1972, reissued 1998); and the commentary of ARISTOPHANES, *Clouds*, ed. by K.J. DOVER (1968, reissued 1989).

Historical background. A view of the religious and political background of the 5th century BC is presented in VICTOR EHRENBERG, *From Solon to Socrates*, 2nd ed. (1973, reprinted 1989); ROBERT PARKER, *Athenian Religion: A History* (1996); and J.W. ROBERTS, *City of Sokrates: An Introduction to Classical Athens*, 2nd ed. (1998). Ancient Sparta is the focus of PAUL CARTLEDGE, *Spartan Reflections*, especially chapters 6 and 7 (2001), pp. 68–90.

The trial. Studies of the trial of Socrates that emphasize both historical and philosophical questions are THOMAS C. BRICKHOUSE and NICHOLAS D. SMITH, *Socrates on Trial* (1989, reprinted with corrections, 1995); and C.D.C. REEVE, *Socrates in the Apology: An Essay on Plato's Apology of Socrates* (1989). Readings from original sources and recent scholarship are presented in THOMAS C. BRICKHOUSE and NICHOLAS D. SMITH (eds.), *The Trial and Execution of Socrates: Sources and Controversies* (2002). JOHN BURNET (ed.), *Plato's Euthyphro, Apology of Socrates and Crito* (1924, reprinted with corrections as *Euthyphro, Apology of Socrates, and Crito*, 1986); and MICHAEL C. STOKES, *Apology of Socrates* (1997), provide line-by-line analyses of the Greek text of Plato's *Apology*. The political setting of the trial is emphasized in MOGENS HERMAN HANSEN, *The Trial of Sokrates—From the Athenian Point of View* (1995). A skeptical view of the historical value of Plato's *Apology* is presented in DONALD MORRISON, "On the Alleged Historical Reliability of Plato's Apology," *Archiv für Geschichte der Philosophie*, 3(82):235–265 (2000).

Athenian persecution of intellectuals. The seminal study is E.R. DODDS, *The Greeks and the Irrational*, chapter 6 (1951, reissued 1973), pp. 179–206. Critical assessments are K.J. DOVER, "The Freedom of the Intellectual in Greek Society," in *Talanta: Proceedings of the Dutch Archaeological and Historical Society*, 7:24–54 (1975); I.F. STONE, *The Trial of Socrates* (1988, reprinted 1994); and ROBERT W. WALLACE, "Private Lives and Public Enemies: Freedom of Thought in Classical Athens," in ALAN L. BOEGEHOLD and ADELE C. SCAFURO (eds.), *Athenian Identity and Civic Religion* (1994), pp. 127–155.

Socrates and religion. Studies of the Socratic conception of piety are M.F. BURNYEAT, "The Impiety of Socrates," in *Ancient Philosophy*, vol. 17, pp. 1–12 (1997); W.R. CONNOR, "The Other 399: Religion and the Trial of Socrates," in *Georgica: Greek Studies in Honour of George Cawkwell*, ed. by MICHAEL A. FLOWER and MARK TOHER (1991), pp. 49–56; MARK L. MCPHERRAN, *The Religion of Socrates* (1996); and NICHOLAS D. SMITH and PAUL B. WOODRUFF (eds.), *Reason and Religion in Socratic Philosophy* (2000).

Early Platonic dialogues. There are several studies devoted to the examination of a single dialogue or a small group of dialogues: R.E. ALLEN, *Plato's "Euthyphro" and the Earlier Theory of Forms* (1970); PLATO, *Gorgias*, trans. by TERENCE IRWIN (1979); W. THOMAS SCHMID, *Plato's Charmides and the Socratic Ideal of Rationality* (1998); PLATO, *Protagoras*, trans. and rev. by C.C.W. TAYLOR (1976, reissued 1996); ROSLYN WEISS, *Socrates Dissatisfied: An Analysis of Plato's Crito* (1998, reissued 2001); PLATO, *Hippias Major*, trans. by PAUL WOODRUFF (1982); and A.D. WOOZLEY, *Law and Obedience: The Arguments of Plato's Crito* (1979). Plato's *Apology* and *Crito* are discussed in R.E. ALLEN, *Socrates and Legal Obligation* (1980). RICHARD KRAUT, *Socrates and the State* (1984), is a study of *Crito* and Socrates' attitude toward politics.

The legacy of Socrates and Athens. Good general overviews are P.J. FITZPATRICK, "The Legacy of Socrates," in BARRY S. GOWER and MICHAEL C. STOKES (eds.), *Socratic Questions: New Essays on the Philosophy of Socrates and its Significance* (1992), pp. 153–208; and C.C.W. TAYLOR, "Socrates and Later Philosophy," in C.C.W. TAYLOR, R.M. HARE, and JONATHAN BARNES, *Greek Philosophers: Socrates, Plato, and Aristotle* (1999), pp. 76–100. Narrower but still valuable studies are A.A. LONG, "Socrates in Hellenistic Philosophy," *Classical Quarterly*, 38:150–171 (1988), and "Socrates and the Sophists," chapter 6 in *The Greek Heritage in Victorian Britain* (1981, reissued 1984), pp. 264–321; and PIERRE HADOT, *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault*, ed. by ARNOLD I. DAVIDSON, trans. by MICHAEL CHASE (1995), pp. 147–178, chapter 5, "The Figure of Socrates."

A history of attitudes toward Athens and Sparta is JENNIFER TOLBERT ROBERTS, *Athens on Trial: The Antidemocratic Tradition in Western Thought* (1994). MELISSA LANE, *Plato's Progeny: How Socrates and Plato Still Captivate the Modern Mind* (2001), is a history of modern debates about the politics of Socrates and Plato. The Socratic transformation of the notion of citizenship and its successors in the modern world is discussed in DANA VILLA, *Socratic Citizenship* (2001). (Ri.Kr.)

Soils

Soil may be defined as the fine earth covering land surfaces that has the important function of serving as a substratum of plant, animal, and human life. Soil acts as a reservoir of nutrients and water and absorbs and oxidizes the injurious waste substances that plant growth accumulates in the rhizosphere (*i.e.*, the root zone). These functions of soil are possible because it contains clay minerals and organic substances (clay and humus form the finer part of soil) that absorb both ions (electrically charged atoms) and water.

This article treats soil formation and the influential factors that are involved, soil profiles, soil properties and their dependence on geographic factors, soil classification, and the geographic distribution of soil. Soil-plant relations are treated in the article BIOSPHERE, THE. For a discussion of the processes of rock weathering and the products that are involved, see GEOMORPHIC PROCESSES and MINERALS AND ROCKS: *Clay minerals*.

For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 232, 353, and 355, and the *Index*.

This article is divided into the following sections:

General considerations 441
Processes of soil formation 441
  Clay formation
  Humification
  The effects of leaching
  Dispersion mechanisms of soil constituents
  Development of soil characteristics
  Rates of soil formation
Basic factors involved in soil formation 445
  Climate
  Drainage and topography
  Vegetation and living organisms
  Parent material
  Time
  Interaction of factors and processes
Soil group classification and nomenclature 447
  Observation of natural categories
  Seventh Approximation System
Geographic distribution of soils 448
  Polar regions
  Podsolic regions
  Brunisolic regions
  Chernozemic regions
  Cinnamonic regions
  Desert regions
  Kaolinitic regions
  Mountainous regions
  Intermediate and modified regions
Bibliography 451

GENERAL CONSIDERATIONS

Regolith and the four horizons

Soil usually consists of a sequence of chemically and biologically differentiated layers, called horizons, that have been formed by the action of natural forces on the unconsolidated residue (regolith) of rocks and minerals on the Earth's surface. The regolith itself is the result of weathering of the original massive rocks at the surface and may be considered to be incipient soil. In the process of weathering, fracturing occurs along planes of weakness in the rocks, and their mineral components react chemically with water, oxygen, and organic, carbonic, nitric, and sulfuric acids that are derived from the atmosphere and living organisms. Through the continuation of these chemical, physical, and biological actions and reactions, the regolith material is transformed into soil that exhibits one to four master horizons called, from the surface downward, A, B, and C horizons (the soil profile) and the underlying consolidated rock, R.

Surface and subsurface horizons are distinguished by the differences in conditions at various depths. Some materials accumulate in the soil, and some substances decompose or are dissociated to release compounds that may be dispersed, dissolved, or transported from one horizon to another; they may even be carried away from the immediate locality. Solutions are present in the soil layers, and these vary in composition and concentrations and thus in the chemical changes they effect. These several conditions determine the abundance and variety of living organisms, from microscopic ones to vertebrates, which in turn influence the further development and the properties of the soil. In the soil profile that results, the horizons differ from each other in characteristics such as colour, chemical composition, particle sizes and distribution, and structure (*i.e.*, arrangement of the particles in groups, aggregates, or independent units).

The A horizon extends from the ground surface to the unmodified regolith or consolidated rock. It is the horizon in which weathering is most intense, with partial removal of the resulting products and a zone of accumulation of organic material. In a typical high-productive soil, the A horizon is rich in humus (more than 1 percent carbon) to a depth of 15 centimetres (6 inches) or more; it contains the greater part of immediately available plant nutrients. Soils of low productive capacity tend to have A horizons that are poor in humus and plant nutrients. The B horizon lies immediately beneath the A horizon and may reach a depth of 65 to 90 centimetres (26 to 35 inches). It is a zone of more moderate weathering in which there is an accumulation of many of the products removed from the A horizon. Sometimes the B horizon is very clayey and impermeable, constituting a serious impediment to plant growth. In a productive soil water, air, and root penetration is easy to a depth of 75 centimetres (30 inches) or more and the water-holding capacity of the entire layer is 15 centimetres (6 inches) of available water, or more. The C horizon contains the parent materials, from which the A and B horizons are formed. The R layer is not part of the soil proper; it is an underlying layer that has properties much different from those of the C horizon immediately above it and that influences the soil by its position.

As a result of soil development factors that will be discussed in following sections of this article, some master horizons can be missing in certain soil profiles. For example, erosion may remove the A horizon; shortness of time or absence of key factors may preclude development of a B horizon; a shallow, weathered layer may be transformed completely into A and B horizons in which case no parent materials are left to constitute a C horizon; or conversely a deep regolith can screen the A layer to the point that the latter does not influence the properties of the overlying soil horizon.

There are many functions provided by soil that are important to human beings. Agriculture and animal husbandry produce more than 90 percent of the human population's food supply; together with forestry these activities connected with soil produce many other materials needed by human beings, such as wood, cellulose, textile fibres, and leather. The utilization of soil by agriculture, animal husbandry, and forestry is often termed "soil exploitation." Soil is necessary for dwellings, highways, airports, and recreation areas, and it also provides road fill and material for water retention structures and fulfills many other essential functions.

PROCESSES OF SOIL FORMATION

The weathering process

Clay formation. The most fundamental soil forming process is weathering, during which crystals of feldspars and other silicate minerals are broken up, releasing chemical compounds such as bases, silica, and oxides of iron and aluminum (specifically the sesquioxides: Fe_2O_3 and Al_2O_3). Leaching (*i.e.,* washing or draining away by percolation of water through the soil) removes most of the bases and some of the silica, and the remaining silica combines with alumina to form crystalline clays.

The kind of crystalline clay produced depends on leaching intensity. Rapid leaching leaves little silica to combine with alumina and results in what are known as 1:1 clays consisting of a tetrahedral (silica) and octahedral (alumina) layer; slow leaching leads to the formation of 2:1 clays, consisting of one octahedral (alumina) layer sandwiched between two tetrahedral (silica) layers. In neither case is the result solely one of the two types, but 1:1 clay is predominant after rapid leaching, and 2:1 clay more abundant when leaching is slow (see Figure 1).

Ions are atoms or groups of atoms that have become electrically charged through the loss or gain of electrons; ions of positive charge are cations and those of negative charge, anions. The phenomenon known as cation exchange is the result of the differences in the force with which clay holds different cations. The surfaces of clay minerals contain many negative charges, which are balanced electrically by the adsorption (*i.e.,* taking up on the surface, as opposed to taking up and distributing throughout the body of an absorbent) of cations. The tightness with which the soil holds cations varies; it is greatest for hydrogen cations and decreases for other cations in the following order: calcium, magnesium, potassium, sodium. Because of the differences in attractive forces, adsorbed cations may be removed and replaced by other cations. Because clay formation is very slow, for long time periods the silica and alumina constitute an amorphous (structureless) clay that has a cation-exchange capacity several times greater than that of crystalline clays. Alumina-rich 1:1 clays have low cation-exchange capacity. Silica-rich 2:1 clays have higher cation-exchange capacity, but lower than that of amorphous clays. Thus young soils and lower horizons that have undergone little weathering have high cation-exchange capacities per unit of clay, as do volcanic soils, because their weathering has been so rapid that clay formation has not had time to take place. Such capacity is lower in old soils in dry or cold climates, and it is lowest in tropical soils.

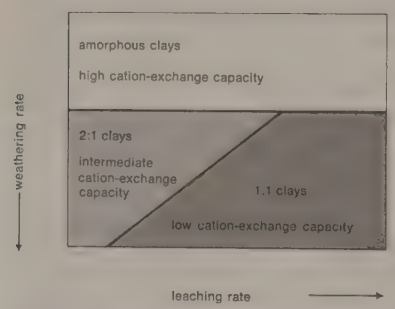


Figure 1: Influence of weathering and leaching rates on clay mineralogy and cation-exchange capacity; rates increase in the direction shown by the arrows (see text).

The most significant aspect of weathering is not its rapidity but its relationship to clay crystallization or leaching. Minerals differ greatly in their weatherability and have been classified accordingly. Clay crystallization seems to be accelerated by a basic medium, especially magnesium. Leaching is an irreversible process (see Figure 2), and thus the effect of a humid season cannot be undone by a dry season. What is important is leaching rainfall, Ln., which is the difference between rainfall and potential evapotranspiration (*i.e.,* return to the air by direct evaporation or by transpiration of vegetation, or both) during the humid season. Leaching rainfall is low in temperate places, even when they are humid (as in London and Moscow), and high in tropical areas, even when they have a long dry season (as in Kano, Nigeria). The prevalence of 1:1 clays in tropical countries is more directly the result of leaching rainfall (see Table 1) than of the rocks' great age.

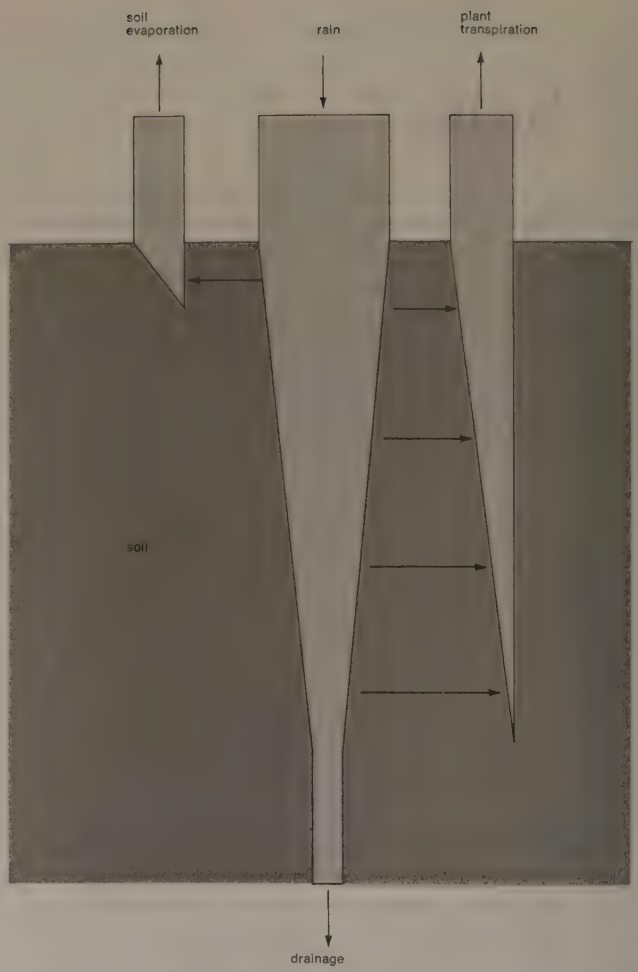


Figure 2: *Water movement in the soil.* In general, water moves in the soil in only one direction—it descends. While descending it is absorbed by plant roots, and the substances it contains precipitate; in this way various substances are transported from the higher layers (horizons) to the lower ones. Under moderate rainfall, the quantity of water that is lost by drainage is little, if any, and only very soluble substances are eliminated.

The character of clay

Rocks subjected to weathering, young soils, and soils formed from easily weatherable materials (*e.g.,* volcanic ashes, lava, or basalt) are rich in amorphous clays of high cation-exchange capacity. Such amorphous clays are mixed with 1:1 clays or 2:1 clays, depending on leaching intensity. Old soils and soils formed from materials that weather with difficulty are poor in amorphous clays and rich in 1:1 clays or 2:1 clays according to whether leaching is rapid or slow. Soil usually contains a great variety of clays that differ in their relative proportions.

In addition to the absorbing capacity for water and ions, which is the most important soil property, clay mineralogy affects such physical properties of soil as its structure and permeability to air and water. A soil rich in amorphous clays has a low apparent density and is very permeable to air and water; decay of organic matter is slow and humus content is unusually high. Soils rich in 1:1 clays usually have a stable structure and swell very little when moistened; on the other hand, a soil rich in 2:1 clays, more especially montmorillonite, has a weak structure, low permeability, and swells easily when moistened.

Tropical soils are usually dominated by 1:1 clays, whereas those of temperate countries contain chiefly 2:1 clays. But pedology, the study of soils, began in temperate countries and is consequently based on what is observed in soils with 2:1 clays; it was some time before the fundamental difference between tropical and temperate soils was real-

ized and understood. For analogous reasons, and because amorphous clays were lost when preparing clay samples for analysis, understanding of volcanic soils has been much delayed. It is only recently that they have been recognized as a special group. Even now the volcanic soils of dry climates are little known.

Mica is a platy silicate mineral that is directly altered (*i.e.*, changed in mineralogical composition) to clays. The first clays formed have high cation-exchange capacity; later 2:1 or 1:1 clays dominate, so that the foregoing outline of the weathering process is applicable to soils formed from materials rich in mica.

Experiments show that weathering is more rapid in a medium containing organic acids. Hence, when soil is covered by raw humus, as in podsolic regions, amorphous clays are formed as a result of the rapid weathering.

By eluviation (the transposition of soil material from one horizon to another) an illuvial horizon is formed by the deposition of materials leached out of an overlaying layer. The illuvial horizon of podsols, called the spodic horizon, is rich in amorphous clays and resembles ando (volcanic ash) soil in many respects.

Clay mineralogy is usually studied by X-rays and techniques of differential thermal analysis, which provide information on structural changes during heating. Soils may be classified on the basis of the relationship between cation-exchange capacity and clay content. Gradations between ando soils and those rich in the clay mineral kaolinite have cation-exchange capacity similar to those of soils dominated by 2:1 clays, but when conditions favour the formation of 2:1 clays, or the soil is young, volcanic soils exhibit very high cation-exchange capacities.

Humification. Second only to clay in importance as a soil constituent is organic matter, which contributes significantly to soil absorbing capacity. Roots die continuously, vegetation and crop residues fall on the soil surface and decay, and the organic leaching products enter the soil. Part of these residues is mixed with soil by organisms living in the soil or by tillage operations; some algae produce organic matter that also is added to soil.

Types of humus

All these residues are transformed by microorganisms into a mixture of organic substances called humus. Two kinds of humus may be distinguished: (1) mild humus is dark in colour, well saturated with bases, especially calcium, rich in humic acids (of high molecular weight), and serves to stabilize clay; (2) raw humus is more red in colour, less basic, rich in fulvic acids (of low molecular weight), and favours dispersion of clay. Soil also contains organic matter that has not yet been humified and in certain cases peat, which is more or less carbonized material that still retains its original structure.

Vegetation determines the kind of humus a soil contains. Grasses produce mild humus rich in humic acids, and conifers produce raw humus rich in fulvic acids. But the base content of soils is also important in determining humus type; for example, calcium-rich soils produce milder humus than do hydrogen-saturated soils. Waterlogging also has an effect, favouring production of raw humus. Root growth is more abundant near the soil surface, and aerial residues accumulate on the surface; thus, organic-matter content decreases with depth.

The effects of leaching. Except for the amount that is immediately evaporated from the surface, rainwater moves through the soil in only the downward direction, and while descending it is absorbed by plant roots. Excess water may be eliminated by drainage, so that even in dry climates the upper layers of soil are somewhat leached. Leaching decreases with depth (see Figure 2). Descending water dissolves various substances, which are precipitated when the water is absorbed by roots. Descending water transports substances from the higher layers to the lower ones, and when some of the water is lost by drainage, the substances it carries are lost as well. More soluble substances are more easily leached away from a layer and are accumulated at greater depth. All chlorides and sodium, potassium, and magnesium sulfates are more soluble than gypsum, which is in turn more soluble than calcium and magnesium bicarbonates.

When there is no drainage, all substances remain in the profile, stratified according to their solubility (see Figure 3)—chlorides and sulfates at the lowest part of the rhizo sphere (root zone of plants), gypsum above them, and calcium carbonate at a higher level. With more leaching the profile deepens, salts are leached away, gypsum accumulates in the lowest part of the rhizosphere, and lime above it. Still further leaching removes first the gypsum, while the lime accumulates in the lower part of the rhizosphere; then it removes the lime, and the soil becomes acidified. Because silica is less soluble than lime, 1:1 clays are usually formed in an acid medium, and are usually acid. But an acid soil may later be basified, and 1:1 clays

Table 1: Leaching Rainfall
(in millimetres)

	Ln		Ln
Bahía Blanca, Argentina	0	Budapest	140
Damascus	10	London	160
Tobolsk, R.S.F.S.R.	10	Berlin	170
Reno	10	Moscow	220
Eureka, Northwest Territories, Canada	20	Jerusalem	230
		Cienfuegos, Cuba	240
Tehrān	20	Stockholm	240
Irkutsk, R.S.F.S.R.	20	Oslo	260
Barrow, Alaska	30	Mexico City	270
Saskatoon, Saskatchewan, Canada	40	Chicago	270
		Rio de Janeiro	280
Baker Lake, Northwest Territories, Canada	40	Amsterdam	310
		Naples	320
Prague, Czechoslovakia	40	Kano, Nigeria	362
Harbin, China	40	Kiel, West Germany	410
Salt Lake City	40	New Orleans	460
Anadyr, R.S.F.S.R.	40	New York	470
Rosario, Argentina	50	Nāgpur, India	517
Nizhnekolymsk, R.S.F.S.R.	50	Bamako, Mali	520
Ankara	50	Kinshasa, Zaire (Congo, Dem. Rep. of the)	640
Omsk, R.S.F.S.R.	50		
Resolute, Northwest Territories, Canada	60	Yaoundé, Cameroon	690
		Djakarta	690
Madrid	60	Wellington, New Zealand	690
Lincoln, Nebraska	60		
Santiago, Chile	70	Canton	820
Corrientes, Argentina	80	Tokyo	870
Athens	80	New Delhi	1,040
Tripoli, Libya	80	Valentia, Ireland	1,150
Tientsin	90	Bombay	1,679
Rabat, Morocco	90	Rangoon	2,000
Niamey, Niger	100	Valdivia, Chile	2,120
Bucharest	100	Cochin, India	2,462
Kiev	110	Conakry, Guinea	3,740
Des Moines, Iowa	120	Monrovia, Liberia	3,860
Paris	130	Cherrapunji, India	10,922
Warsaw	130		

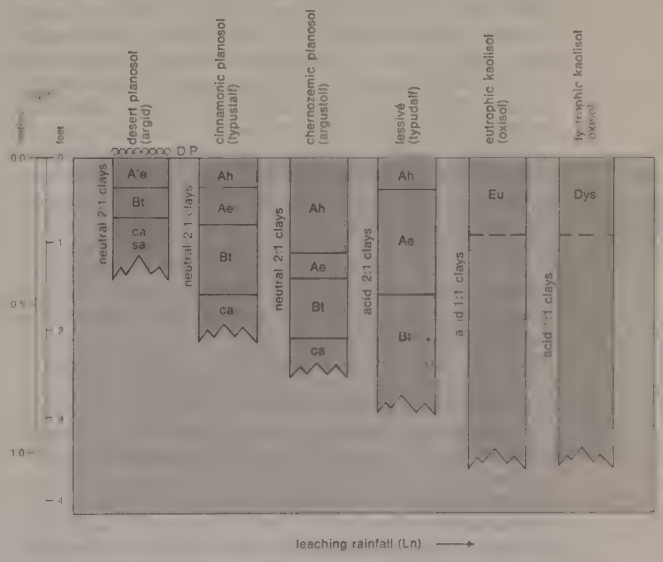


Figure 3: *Influence of leaching rainfall (Ln) on soil formation.* Leaching rainfall increases from left to right, as shown by the arrow. D.P., desert pavement; A'e, surface horizon poor in humus and eluvial; Bt, textural B; ca, lime accumulation; sa, salts accumulation; Ah, surface horizon rich in humus; Ae, eluvial surface horizon; Eu, eutrophic upper soil; Dys, dystrophic upper soil (see text).

are sometimes inherited from parent material, as in the case of allochthonous soils (*i.e.,* those transported from other environments).

Dispersion mechanisms of soil constituents. *Biological transport of elements.* Plants absorb various elements from the profile and deposit them, with their residues, on the soil surface, a process that counteracts leaching. In soils rich in bases leaching is more important. This accounts for the fact that base saturation and pH (the standard measure of acidity, related to the concentration of hydrogen ions; a pH value of 7.0 is neutral, whereas that from 1.0 to 7.0 is acid and that from 7.0 to 14.0 is alkaline) usually increase with depth. In acid soils, leaching is difficult because cations are firmly retained by soil colloids, and biological transport to the surface prevails; base saturation decreases with depth. In the kaolisols of the humid tropics, the upper 25 centimetres (10 inches) of soil often contain more bases than the rest of the profile to a depth of 150 centimetres (59 inches). The process is selective, however; *e.g.,* sodium and magnesium are less absorbed by plants than are potassium and calcium.

Base saturation and pH

Addition of constituents by running water and wind. Floodwaters always contain salt, and soil, when flooded, may become saline (*i.e.,* high in chlorides and sulfates of sodium, potassium, calcium, and magnesium) unless some of the water is lost by drainage. Floods sometimes deposit fine earth or coarse materials on the soil surface. Sea winds often bring salts, and some soils, such as those of New Zealand and the Falkland Islands (Islas Malvinas), have probably been salinized in this way. In other cases winds bring dust or volcanic ash, additions that may counteract leaching and modify the soil profile.

Subirrigation. In some soils there is a water table one to two metres below the rhizosphere. During the dry season water rises by capillary action from this water table to the rhizosphere, where it is absorbed by plant roots, and the substances it contains accumulate. In this way a horizon of salts, gypsum, lime, silica, or iron accumulation is formed, depending upon the substances that are contained in subirrigation water. When the accumulation takes place in a layer poor in clay, the lime, silica, or iron are deposited on the surface of sand or gravel and may cement them, forming a hardpan (cemented layer impenetrable by roots). Many calcareous horizons, duripans, fragipans, and iron pans have this origin, but they also may be cemented by substances that descend in the profile. When the water table is shallow, the hardpan may reach the soil surface.

Solonization. Sodium clay is easily dispersible, and in sodium soils it is transported from the surface to some depth. Sodium soils swell when moistened, however, which impedes clay movement, so that clay illuviation takes place near the surface rather than at depth. This process is called solonization, or solonetization, and results in sodium-saturated soils (solonetz) that have a richer clay horizon at depth than at the surface.

Leaching activated by vegetation, however, produces organic acids and transports calcium from the lower layers to the surface. Sodium then is replaced by calcium or hydrogen in the eluvial layer and the upper part of the illuvial one. Clay illuviation thus takes place at a greater depth, and soil becomes planosol or solod (calcium- or hydrogen-saturated soils with considerable difference in clay content between the eluviated surface and the clay illuvial horizon).

Transport of chemicals in soils

Planosols can also be formed directly from a slightly sodium-saturated soil; low sodium saturation is sufficient to facilitate clay dispersion and permits clay movement to greater depth. Solonetz are often formed from sodium-saline soils, and planosols from easily weatherable materials, such as volcanic ash, under conditions that impede leaching (*e.g.,* dry climate or poor drainage or both).

Iron and clay illuviation. Fulvic acids react with iron and may cause its eluviation and that of clay. For that reason, soils formed under forest vegetation, which produces humus rich in fulvic acids, are often leached; there is a difference in clay content between the surface horizon and the B horizon in which the clay is illuviated. Moreover, such soils are relatively rich in free iron, and a difference in free iron content is often observed between the surface horizon, which is poorer, and a lower enriched horizon. Because dispersed iron is easily precipitated by bases or concentration of soil solution, soils developed under conditions that favour the formation of fulvic acid-rich humus are usually rich in iron that is finely precipitated with organic matter, which gives them a brown colour (brunisolic soils).

Podsolization. Weathering in a strongly acid medium produces an eluviation of the sesquioxides (Fe_2O_3 and Al_2O_3) released by weathering. Because weathering is very rapid amorphous clays of high cation-exchange capacity are formed. The soil formed has an upper horizon rich in unweathered minerals (sand), and an illuvial one rich in amorphous clays ("spodic" horizon). Such soils are called podsols. Although vegetation and other conditions may favour the accumulation of raw humus, its acidity is neutralized when soil contains lime or when weathering releases bases in sufficient quantity. Podsolic weathering is therefore observed in sandy materials rich in quartz and other minerals of low weatherability. The organic matter that produces eluviation of clay or sesquioxides decays and cannot accumulate in the illuvial horizon; but when this horizon is waterlogged or when its temperature is very low, organic matter accumulates and a humus illuvial horizon forms. Such soils are called humus podsols.

Rubification. When soil is thoroughly dried from time to time, precipitates of iron and organic matter cannot accumulate and the organic matter disappears by decay, causing irreversibly dehydrated iron sesquioxides to form. This process is known as rubification, and the resulting soils, called cinnamonic, are rich in fine crystallites of dehydrated iron sesquioxides.

Ferrugination. Iron sesquioxides adhere firmly to sand grains and gravel, give them a red colour, and may cement them to form an iron pan. This process is most common in soils that have been formed from materials that release much iron upon weathering and leave much sand, whereas humic horizons are seldom ferruginized, probably because organic acids and reductive processes remove the iron coating from sand grains. Fulvic acids favour iron eluviation and the formation of ferruginized horizons at some depth; thus, fragipans and other iron-cemented pans are common in podsols and brunisolic soils. The ferruginous horizons of tropical soils are formed similarly, fulvic acids contributing to their formation, but reductive processes can also be responsible for iron eluviation. Periodic soil drying may consolidate iron coatings, especially in climates in which a humid and a dry season alternate. It seems that iron coatings protect sand and gravel grains from weathering, and they are the cause of many "stone lines" parallel to soil surface.

Formation of cemented pans

Segregation of iron. In soil that is rich in iron and poor in 2:1 clays that retain it, iron tends to segregate, forming irregular concretions and nodules. After repeated moistenings and dryings, the nodules become as hard as stone and may form a pan cemented with iron and gravel (laterite). Waterlogging seems to further iron segregation, and alternative moistenings and dryings harden the nodules or pan. This process resembles ferrugination, the difference being in degree of segregation.

Gleization. Under waterlogging conditions, iron becomes ferrous and has a valence of 2+. (Valence is an expression of the combining power of an element, which is determined by the number of electrons in its outer or valence shell of electrons; valence is represented by a number that corresponds to the number of electrons an atom of the element can accept [if the number is positive] or donate [if the number is negative]). Ferrous soil has gray-blue colours, is called gley, and the process is called gleization. When waterlogging is less severe and transitory, ferrous iron is oxidized (*i.e.,* loses an electron, with a resulting change in valence to 3+, which is that of ferric iron, F^{3+}), and ruggy mottles (spots) are produced in the soil. Because ferrous iron is more mobile than ferric iron, gleization favours iron eluviation, resulting in a leached upper horizon; iron accumulates at a lower depth or is leached away. Or the iron may ascend by capillarity from the water table to the lower part of the rhizosphere.

Development of soil characteristics. *Soil profile and*

horizons. Humification, leaching, clay eluviation, and other soil-forming processes all create differences between soil horizons. Soil profiles vary considerably in horizon types and their thicknesses.

Soil properties. Both water-holding capacity and cation exchange depend on clay content, clay mineralogy, humus content, kind of humus, and nature of absorbed ions. Mineral constituents larger than clay also have some water-holding capacity, and in some cases silt exhibits considerable ion exchange capacity. This may be attributed to its formation by aggregates of amorphous clays or by amorphous clay that adheres to silt particles.

Nature of soil texture and structure

Soil texture (*i.e.,* the relative proportion of mineral constituents classified according to their size) determines to a considerable extent soil porosity and, consequently, permeability to air and water, which are essential for root life and many processes that take place in soil. A high content of coarse constituents usually increases porosity. But clay and humus bind single soil grains together in increasingly larger aggregates.

Soil has a "structure" on which its porosity-permeability depends greatly. Soil structure is built up by alternate moistening and drying, and plant roots contribute greatly by opening pores between soil aggregates; it is undone by waterlogging. The stability of aggregates increases with humus content, especially humus that originates from grass vegetation (forest humus is less effective).

Calcium and hydrogen soils have structural stability, whereas sodium soils lose stability easily. Soils rich in amorphous clays are very porous and permeable, and their apparent density is exceptionally low. Soils rich in 1:1 clays are also permeable. High content of 2:1 clays, however, is usually associated with low permeability because the soil swells when moistened and the pores close. The clay type is very important; montmorillonite is the worst in this respect. Water movement in 2:1 clayey soils is chiefly through cracks produced when soil dries, and water movement in pores between small aggregates is very slow. Soils rich in coarse constituents are usually permeable. Soils with a high fine-sand content, however, have low permeability, poor aggregation, and small pores between single grains of fine sand; moreover, they are not extensible, and they resist root penetration.

Texture variations in horizons

Because soil horizons vary in texture, humus content, and absorbed ions, soil properties vary from horizon to horizon. This fact should be taken into consideration in evaluating the soil as a whole. Distinction must be made between intrinsic qualities—such as texture, clay mineralogy, the nature of absorbed cations, and their variation along the profile—and the secondary qualities that are consequences of the first and are more or less transitory—such as structure and permeability. Humus content and soluble salts are intermediary; they can be modified by management, added, or leached away.

From an agricultural point of view, the most important soil qualities are cation-exchange and water-holding capacities, the nature of absorbed cations, richness in assimilable plant nutrients and certain other substances (*e.g.,* lime, free aluminum), permeability, structure, humus content, and waterlogging. For engineering uses, texture, clay mineralogy, humus content, swelling capacity, capacity to corrode metals, structure (which can be modified), permeability, and waterlogging (chiefly extrinsic), are important.

Horizon nomenclature. Usually the upper horizon, which is generally richer in humus and may be eluvial, is called A; it may be "humic" (Ah), eluvial (Ae), or "bleached" (A_2). The horizon in which clay has been illuviated is called Bt; that of amorphous clays produced by podsolization ("spodic" horizon) Bir (from iron, which was considered the most important illuviated substance); and that of humus illuviation Bh. Weathered material that has not been sufficiently enriched in humus to be humic and has not had important illuviation of clay or humus is called C. Underlying consolidated rock is called an R horizon.

A horizon of lime accumulation is usually called ca, but the term can be extended to all horizons that give effervescence with an acid. An accumulation of gypsum is indicated as cs; a horizon unusually rich in easily soluble salts by sa; cn designates an accumulation of ferruginous concretions and nodules; mf a ferruginous hardpan; mc a calcic hardpan; x a fragipan (fragil hardpan cemented with iron); G denotes strong gleying; g moderate gleying.

According to clay mineralogy a horizon may be allophanic, illitic, kaolinitic, or superkaolinitic, the diagnosis being chiefly based on the relation between cation-exchange capability (CEC) and clay content. According to base saturation, a horizon may be acid, neutral, natric, magnesic, or calcic. Fifteen percent sodium saturation is sufficient to make a horizon "natric." The terms acid and neutral are replaced by dystrophic and eutrophic (poor or rich in nutrients) in the case of kaolinitic and organic soils in which the relation of absorbed cations to clay content is a better diagnostic indicator than base saturation. Many of these terms are poorly defined in the literature, and definitions vary from author to author.

Rates of soil formation. Rates of soil formation vary enormously according to the process involved, and for the same process according to conditions. Weathering varies according to the weatherability of the material. and minerals may be classified in this respect. Weathering of a consolidated rock is usually slow at the beginning, accelerates as the rock breaks up and more surface area is exposed to attack, and then becomes increasingly slower as weatherable minerals disappear. It is accelerated by plant roots and by the accumulation of raw humus, and also by a rapid removal of the bases by leaching.

Mineral composition and rock weathering

Weathering is more rapid than is usually thought. In experimental leaching of coarse gravel of basalt with pure water at 20° C (68° F), 0.45 percent of the silica was eliminated in 11 months. Leaching with water containing acetic acid (pH 2.5, which is a common acidity value in the vicinity of roots and decaying organic matter) accelerated the process by a factor of 21. Volcanic ashes weather sufficiently to support vigorous vegetation in a few decades.

Tombstones and other man-made structures weather more slowly because they are not in contact with plant roots or decaying organic matter, and, moreover, the surface of attack is very small.

Humification is a very rapid process. Experiments and agricultural practice all over the world have shown that soil organic matter can be doubled or reduced to half in 5–50 years. Podsolization is also rapid, and it has been reported that soils planted with conifers have been podsolized in a century. Salinization and desalination are also rapid because of the higher solubility of salts less soluble than gypsum. Decalcification and acidification with water containing carbon dioxide are intrinsically slower but may be accelerated when produced by water containing organic acids, which is usually the case.

There is less information concerning clay eluviation, but experiments show that this is a relatively rapid process because water often moves through soil cracks, carrying much clay per unit of water. The formation of argillic (clay illuvial) horizons has been estimated to require between 550 and 5,250 years. In many cases the age of soil that has been formed in previously glaciated areas, newly formed terraces, or from recent volcanic material, is known; but it is not always possible to ascertain whether clay eluviation required all this time. In any case, existing evidence and the fact that soils correspond well to present climatic conditions indicate that soil formation processes are relatively rapid.

BASIC FACTORS INVOLVED IN SOIL FORMATION

Climate. The amount and kind of clay that is formed, soil depth, base saturation, and the formation and depth of saline, gypsic, or calcic horizons all depend on leaching intensity. Consequently, the significant factor is leaching rainfall (see Figure 3). The soils of temperate climates are rich in 2:1 clays, whereas those of tropical countries are rich in 1:1 clays. But in both cases when the soil is young, or has been formed from volcanic ashes, it contains considerable amorphous clays, which increase its cation-exchange capacity.

Temperature is another influence in soil formation. Because the solubility of silica increases with temperature,

high temperatures favour the formation of 1:1 clays, which is another reason for the abundance of such clays in tropical soils. Water penetration and root penetration depend on leaching rainfall, so that the depth of weathering and of soil increase with Ln. Soils of tropical countries are usually very deep; those of cold or dry climates, or of both, shallower. Temperature is also important. At high latitudes and altitudes the lower layers of soil are very cold, even in summer; consequently, weathering and soil depth are limited, which is important in the case of autochthonous soils (*i.e.,* formed in the original location) developed from consolidated rocks. Permafrost limits both soil depth and drainage. In very dry climates no water is lost by drainage. In dry climates substances formed by weathering accumulate in the profile, stratified according to their solubility, and as leaching increases the lower horizons disappear, and the profile becomes deeper. Finally, there are insufficient bases in the soil to neutralize the organic acids produced by decay of roots and organic matter. Solution in the soil then tends toward acidity, the soil is leached with acid water, and hydrogen ions replace other ions. The result of this cycle is an increasingly more acid soil. Acidification, however, also depends greatly on vegetation type, temperature, and waterlogging. In addition to the above direct influences, climate greatly affects vegetation, which is one of the most important factors in soil formation.

Effects of Ln on weathering depth

Drainage and topography. Leaching is impossible without drainage; thus, impeded drainage has to some extent the same effect as drought. Impeded drainage may result in the formation of soils with 2:1 clays, well saturated with bases, saline, gypsic, or calcareous horizons. It also causes gleization and may result in the formation of a water table, which has various influences, especially salinization and formation of various pans—petrocalcic horizons, duripan, ironstone, fragipan, laterite, etc. (see Figure 4). Topography greatly influences erosion, drainage, flooding, and subirrigation, which have a profound influence on soil formation, as has been discussed. Soils on slopes are usually shallow and stony (lithosols), and those in the valleys are alluvial.

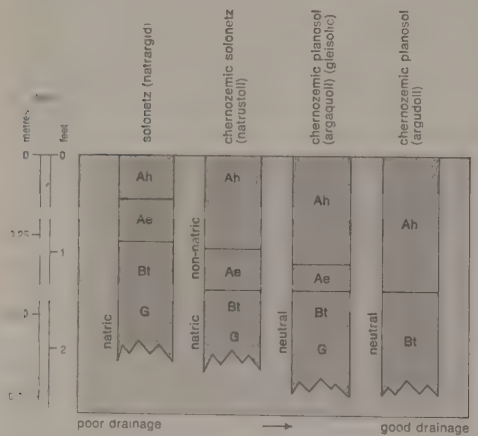


Figure 4: *Influence of drainage on soil formation.* Ah, rich in humus surface horizon (layer); Ae, eluvial surface horizon; Bt, textural B (clay illuviation); G, gley.

Variation in humus

Vegetation and living organisms. Vegetation largely determines the kind of humus that is formed. Some trees, especially conifers, usually produce raw humus that may cause podsolization. Other trees produce weakly acid humus capable only of producing iron mobilization and clay illuviation. Such soils are brown (brunisolic) and more or less leached. Grasses produce mild humus rich in acids that stabilize clay. The distribution of humus along the profile also depends on the type of vegetation. Grasses have numerous roots that decay rapidly and enrich soil in humus to considerable depth. Tree roots contribute little to soil humus; thus, the humic horizon of forest soils is usually shallow (see Figure 5).

Vegetation protects soil from the Sun's rays and retards organic matter decay; it absorbs calcium, potassium, and

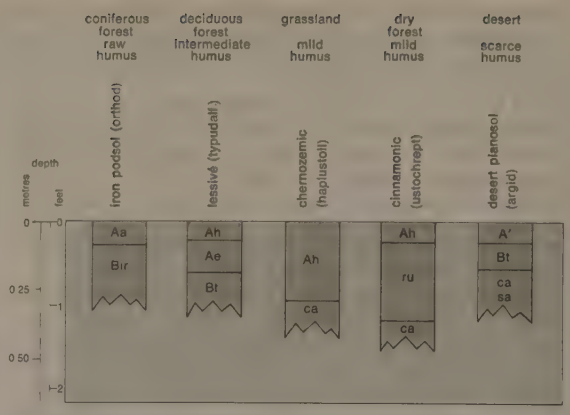


Figure 5: *Influence of vegetation and humus on soil formation.* Aa, albic (ashy) surface horizon (layer); Bir, iron podsolic B (spodic horizon); Ah, humus rich surface horizon; Ae, eluvial surface horizon; Bt, textural B (clay illuvial); ca, lime accumulation; ru, rubified, A', poor in humus surface horizon; sa, salts accumulation.

phosphorus from the lower soil layers and counteracts their leaching; and it protects soil against erosion. Erosion, even hydraulic, is severe in deserts because there is no vegetation to protect the soil. Living creatures within the soil such as earthworms and termites mix the various soil horizons and counteract their differentiation; they also mix soil plant residues on the soil surface and contribute to organic matter decay and humus formation. A certain mixing also occurs when old trees are uprooted by wind.

Parent material. Soil is parent material that has been altered by various processes, and its ultimate nature depends upon its original composition. In soils formed from unconsolidated sedimentary rocks, much of the clay is inherited from parent material. Materials rich in lime are difficult to decalcify, acidify, or podsolize. Materials rich in iron usually give red soils.

Time. Although in geological terms soil formation is rapid, in human terms it requires many decades or centuries. Soil age has a great influence on soil properties. The age of a soil usually is measured from the time the rock was first exposed to the surface or from the time the alluvial material was deposited. The time factor is not always clear because erosion and deposition are continual processes. Moreover, clay is often much older than the deposition, having been formed in one place and transported to another.

Young soils commonly have an undifferentiated profile. Saline, gypsic, and calcic horizons are formed more rapidly than those of clay accumulation, whereas a humic horizon may be formed in a few decades. Young soils are richer in amorphous clays and, under conditions that favour acidification, are less acid than old ones; pH increases with depth in young soils, and it decreases with depth in old ones.

Interaction of factors and processes. Any effect that each factor has on soil formation depends on the other factors; for example, the effect of climate depends on drainage, parent material, time, etc., and the other factors are similarly interdependent. Different soils are formed under the same climatic conditions according to drainage, parent material, time, vegetation, and other relevant influences; in like manner, different soils may be formed from the same parent material. Because soil is the result of various processes and each factor acts on several of them, it is better to relate soils to processes, and through processes to factors. The environment under which a soil is formed may vary during its formation: drainage may improve or deteriorate; exposure to such factors as erosion, flooding, or subirrigation may be ended or initiated; or there could be a change in vegetation. Thus soil does not always reflect the conditions under which it is encountered.

The polygenesis of soils

Climatic changes are very slow, and to affect soil formation appreciably they must be radical. A change in drainage will impede leaching, just as will a very drastic reduction of leaching rainfall; waterlogging may cause the formation of raw humus that produces podsolization in a

warm climate, that otherwise favours organic matter decay and excludes podsolization, when drainage is normal. Little leaching rainfall is needed to leach a coarse soil; a great deal of rainfall is needed to leach a clayey soil. Acidification is faster and more extensive when parent material is poor in bases. For all these reasons, caution should be used when attributing the properties of a soil to climatic changes or, still more, when explaining climatic changes on the basis of pedologic evidence.

There are cases, however, in which such evidence is convincing. In Australia and Africa, for example, soils that are rich in 1:1 clays are found in the desert; they have most certainly been formed under a considerably more rainy climate. Because allitic weathering (that producing chiefly 1:1 clays) is deep, subsequent erosion and deposition merely transport such soils from one place to another. In some cases 1:1 clays have been transported great distances from where they were formed: for example, in the delta of the Paraná, near Buenos Aires, there are soils with clays that came from the centre of Brazil, and many soils of northern India have 2:1 clays that have been formed in the Himalayas.

The frequent variation in conditions during soil formation, resulting in polygenetic soils, usually is more closely related to topography than to climate; also, the clay found in a soil may have been formed in past time or at a distant site.

Table 2: Major Soil Types

Ando
rich in amorphous clays

Brunisolic
2:1 clays; braunified or acid horizons

Chernozemic
2:1 clays; deep neutral dark humic horizons

Cinnamonic
2:1 clays; reddish colour; rich in dehydrated iron oxides

Dark clays
2:1 clays; rich in expanding clays; well saturated with bases

Gleisolic
gray-blue colours or mottles due to waterlogging

Kaolinitic
tropical; rich in 1:1 clays; may have undifferentiated profile, or horizons rich in iron concretions or laterite

Lessivé
clay illuviation produced by organic matter

Organic
extremely rich in organic matter; may have peaty and, or organic horizon

Podsols
eluviation of alumina and iron, usually producing ashy sandy surface horizon and illuvial horizon rich in amorphous clays

Planosols
clay eluviation produced by sodium; illuvial horizon not natric

Rankers
2:1 clays; humic horizon resting on rock or permafrost

Raw
2:1 clays; undifferentiated soil profile

Rendzinas
high lime content

Solonchaks
rich in salts such as chlorides, sulfate, etc.; related to gypsisols, rich in gypsum

Solonetz
clay illuviation produced by sodium; natric illuvial horizon

SOIL GROUP CLASSIFICATION AND NOMENCLATURE

Observation of natural categories. The question of soil classification is a controversial one, but there is no doubt that classification should be based on categories that exist in nature, rather than on arbitrary creations of a classification system.

Rationale of group selection

In a natural classification, groups are recognized and subsequently arranged in a system; the system may change, but the groups will pass almost unchanged from one system to another. Pedologists have approached soil classification in such a manner. Observing that some soils have an ashy surface horizon, resting on another darker in colour and richer in fine earth, pedologists called them podsols (ashy soils). Other soils with a deep, dark horizon rich in organic matter and well saturated with calcium were termed chernozems (black soils). Soils of rather undifferentiated profile and of red colour were named krasnozems (red soils). In this way, groups of soils have been recognized and arranged in classification systems; although opinions differ as to how these groups should be arranged within a classification system, essentially the same groups exist for all pedologists.

The soil continuum. Because soils form a continuum, it has been difficult to follow the example of other scientists in categorizing on the basis of clear distinctions between groups. The same processes are involved in the formation of various soils, with the result that a soil may have the essential features of ando soils (amorphous clays) and those of chernozemic soils (deep humic horizon, dark and well saturated with bases) and thus be both ando and chernozemic. Soils may reflect many such heterogeneous conditions—for example, chernozemic (mollic) and gleisolic, gleisolic and podsol, and the like. Accordingly, soil classification must be manifold, recognizing that the same soil may belong to various groups. When this necessity is understood, the problem of soil classification is simplified, and the major categories shown in Table 2 may be recognized (see Figure 6).

Fundamental classes of soil

Supplementary descriptions. Various other terms, denoting special features, also may be precisely and simply defined and included in soil nomenclature. For example: clay, loam, sand (referring to soil texture); lithosol (stony, gravelly); regosol (sandy); aeolian (formed by wind, dune); hammada (stony desert); and rock outcrop. Acid (H s.), natric (Na s.), magnesic (Mg s.), Ca soil, eutrophic, and dystrophic all refer to absorbed cations. Takyr and lunettes are special kinds of saline soils. Roof clays and sulfate saline clays (cat clays) are special kinds of clays. Gray-brown podsolic, noncalcic brown, graywooded, red-yellow podsolic, and lateritic podsolic are different kinds of lessivé soils; the first three have 2:1 clays, the later two are kaolinitic. Sod and humic are used for soils unusually rich in humus for their groups. Prairie, chestnut (kastanozems), and reddish chernozemic are different kinds of chernozemic soils. Podsolized, leached, or degraded chernozems are types of brunisolic or leached soils or both. Claypan planosols are planosolic clays. Serozems are rendzinas lacking humic horizon. Rubrozems are kaolisols with a deep humic horizon. Red desert are mini-planosols (thin horizons). Mini-planosols, mini-solonetz, and mini-podsols are planosols, solonetz, or podsols with thin horizons. Arctic browns are brunisolic permafrost rankers. Terra rossa is a cinnamonic soil from hard limestone. Terra roxa is an ando-kaolisol intergrade soil. Terre de barre is a eutrophic latosolic kaolisol. Low humic gley is an acid gleisolic soil. Humic gley is a gleisolic chermozemic soil. Desert crust is an accumulation of products of weathering in virtual absence of leaching. Alluvial, desert, forest, forest-steppe, grassland, groundwater, lowland, mangrove, meadow, mountain, paddy, paramo, prairie, tropical, and tundra are environmental, nontaxonomic terms often used in soil nomenclature.

Precise definition of soil properties

Seventh Approximation System. A system based on measurable soil properties rather than on theories of soil formation, the 7th American Approximation System, was introduced in 1960 and has expanded continuously by supplements. It is the first approach to precise definition of soil groups and pedologic terms, and, moreover, it has implicitly adopted manifold classification: adjectives (mollic, spodic, andic, etc.) are used to show relationships with other groups; names are formed by aggregation of formative elements that denote special characteristics, and consequently groups. Many of the groups named in the preceding section correspond in 7th Approximation to several taxa the nomenclature of which includes the corresponding formative element. For example, andepts and andic correspond to ando; histosols to organic; spodosols to podsols; vertisols to dark clays. The formative element aqu denotes gleisolic; natrarg, solonetz; rend, rendzina; sal, saline.

Indication of soil properties. Some formative elements show soil properties: acr denotes extreme weathering; allic, free aluminum; anthr, man-made soil; arenic, sandy; calci, calcic; chrom, high chroma (red colours); dys or dystr, dystrophic; eu or eutr, eutrophic; gibbs, presence of gibbsite; hydr, saturated with water; lithic, stony; pale, old (soil); pell, dusky colour; psam, sandy; rhod, red; sal, saline; sider, rich in free iron; sombri, dark colour; thapto, a buried soil is included in the profile; vermi, wormholes, worm casts, or filled animal burrows; vitr, ando rich in unweathered volcanic glass.

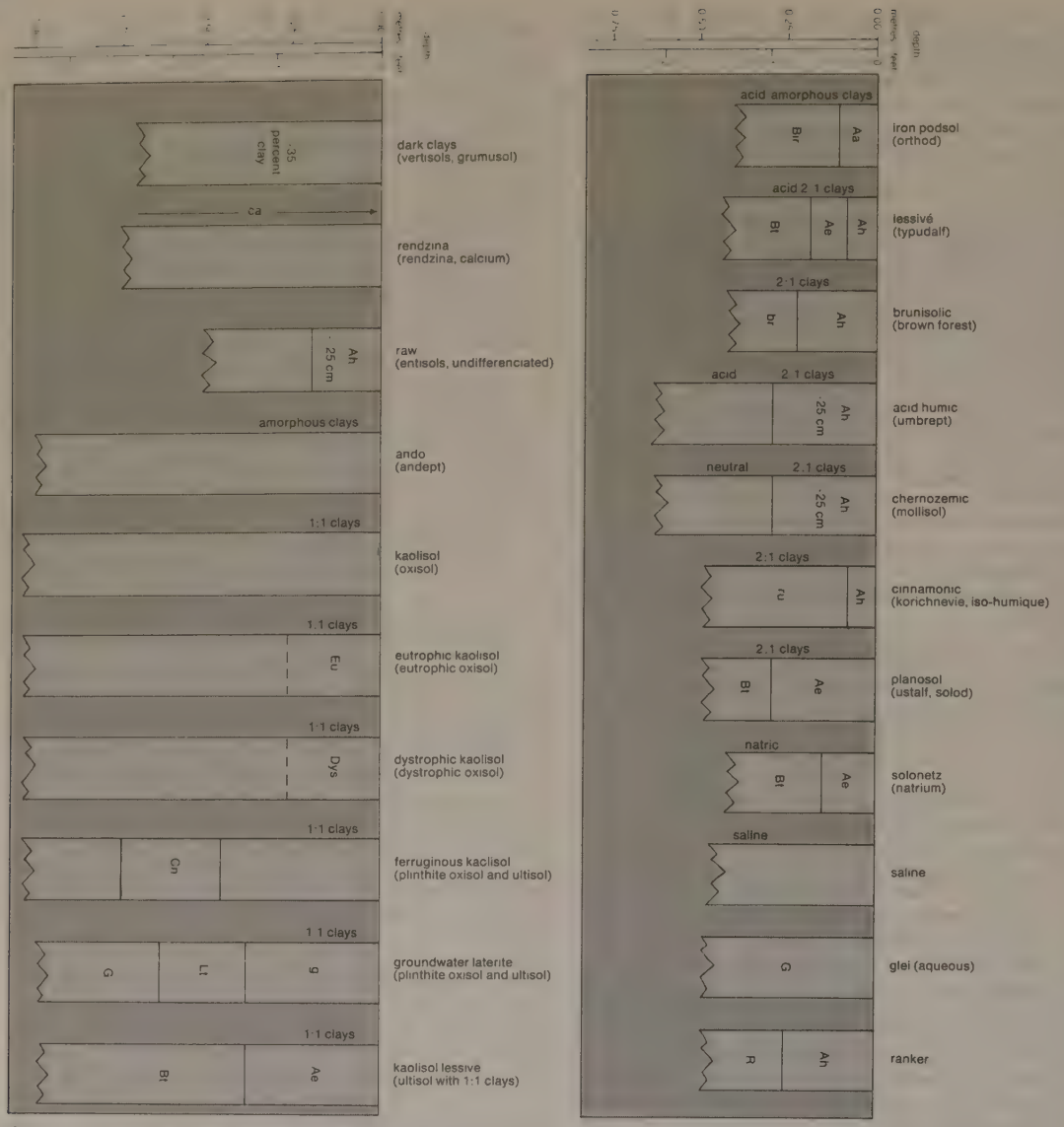


Figure 6: *Fundamental characteristics of some soil groups.*
Specifications referring to reaction (acid, neutral) and to type of clay refer to the whole profile. Aa, albic (ashy) surface; Bir, iron podsolic B (spodic); Ah, humus rich surface h.; Ae, eluvial surface h.; Bt, textural (clay illuvial) B; br, braunified; ru, rubified; G, gley; R, rockbed; Eu, eutrophic (rather well provided with bases) soil; Cn, ferruginous concretions and nodules; Lt, laterite; Dys, dystrophic (poor in bases) upper soil.

Identification of horizons. Other formative elements denote presence of certain horizons: abruptic, abrupt textural change; alb, albic horizon; cumulic, thickened humic horizon; dur, duripan; ferr, B horizon rich in free iron, or iron concretions or nodules; frag, fragipan; glossic or gloss, horizon A_2 is tongued; hapl, typical profile; hum, unusually deep humic horizon, or humus podsolic B horizon; ochr, light-coloured A horizon; pachic, thick A; petrocalcic, petrocalcic horizon; plac, thin pan; plinth, laterite; ruptic, broken (discontinuous) horizon; stratic, stratified horizon; superic, laterite in the surface; umbr, deep nonchernozemic humic horizon. Other formative elements of the nomenclature indicate the manner in which the material has been deposited: fluv; alluvial; limnic, lacustrine.

Climatic elements of the system. Although soil temperature and moisture are not intrinsic soil properties, they are used as diagnostic features in the 7th Approximation; *i.e.*, soils are classified according to the climate in which they are encountered. Such formative elements include: bor, boreal; cryo, very cold; pergelic, permafrost; torri, torrid (usually dry); trop, small annual range of temperature; ud, usually humid; ust, dry but not long dry seasons; xer, long dry seasons.

The 7th Approximation System demonstrates that soil description, classification, and nomenclature may be achieved with few symbols that, in combination, can represent the enormous variety of soils existing in nature.

GEOGRAPHIC DISTRIBUTION OF SOILS

Because soil formation depends so much on climate, topography, vegetation, and parent material, soils vary in association with other geographic features, and soil regions may be distinguished in the same manner as climatic and vegetation regions. Such soil regions lend themselves to being further subdivided or combined into higher groups. Each region contains a variety of soils, many of which are also encountered in other regions, but the distribution of soils follows a definite pattern.

The concept of soil region is important from both a theoretical and a practical point of view. Theoretically, soil distribution is often the key to answering many questions about soils. From a practical viewpoint, it is difficult and costly to prepare detailed maps of soil locations, which can vary in short distances—sometimes within a few metres.

The function of soil maps

Polar regions. *Tundra.* Vegetation is scarce in tundra regions (see Figure 7) and consists of herbs and grasses that do not produce very acid humus. Rainfall in the tundra is

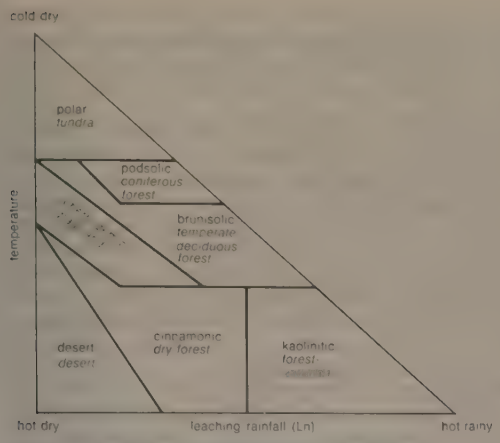


Figure 7: Distribution of soil regions with associated vegetation according to temperature and leaching rainfall (see text).

low and leaching rainfall (Ln) is still lower: 40 millimetres (one inch equals 25 millimetres) in Baker Lake, Northwest Territories; 20 in Eureka, Alaska; 60 in Resolute, Northwest Territories, Canada; 30 in Barrow, Alaska; 40 in Anadyr, 70 in Bulun, 50 in Nizhnekolymsk, and 40 in Russkoye Ustye, Siberia. The bases that are leached from the surface accumulate in the lower part of the profile and circulate in it through plants and by capillarity (the soil surface is often humid and soils are shallow); moreover, drainage is often impeded by permafrost. Acidity is not sufficient to produce podsolization, but it reacts with iron to produce complexes that result in brunisolic soils. The soils formed, arctic brown, are relatively shallow; the pH increase with depth is more rapid than usual; and there is a tendency toward the formation of a humus-rich horizon on the surface of permafrost. In the drier parts of tundra, the lower horizons are often calcareous.

Waterlogging is frequent in tundra because of the presence of permafrost, resulting in the accumulation of organic matter, a prevalence of reducing conditions and peat formation. Gleisolic and organic (half bog) soils are frequent. The general pattern of distribution is arctic brown in well-drained sites to gleisolic and half-bog mini-podzols in depressions. Alternate freezing and thawing results in a patterned surface (polygonal).

Subglacial desert. In the subglacial desert vegetation is almost absent, weathering is very slow, and the soils formed are very shallow (lithosols, etc.). The limits of tundra and subglacial desert soil regions correspond rather well with the homonymous climatic regions.

Podsolic regions. With a short summer and conditions not too frosty for forest, so that an evergreen coniferous forest prevails, a raw, highly acid humus is produced and podsolization takes place. The soils formed in well-drained sites are iron podsols.

Drained and waterlogged conditions

In waterlogged sites organic matter accumulates, reducing conditions prevail, and gleisolic soils, sometimes organic, are formed. When there is no drainage, bases cannot be eliminated from the profile and soil is less acid, a condition common in the drier podsolic regions. When there is a slow drainage, however, soil is acidified and becomes humus podsol, somewhat gleisolic and peaty. Such soils usually have heath or sphagnum vegetation.

Humus podsols are more common in Atlantic regions because of their higher rainfall, than in continental regions. Hardpans are more common in Atlantic regions. Some regions of Central Siberia have permafrost that interferes with drainage; moreover, leaching rainfall is low, with the result that iron leaching is incomplete and the eluvial horizon is yellow rather than ashy. The yellow podsols may be considered as intergrades to arctic brown.

Brunisolic regions. With a longer summer, deciduous forest prevails, which produces milder, less acid humus that forms brunisolic soil, more or less leached. Where leaching rainfall is low and parent materials are calcareous or basalt, soils are neutral (braun erde); in the opposite rainy condition, soil is acid (brun acide).

The most important brunisolic regions are in western Europe, the eastern United States, the state of Washington, and the adjacent Canadian coast. In western Europe, there are many small regions with calcareous materials, basalt, or loess, and low leaching rainfall. Braun erde abound; some of them are chernozemic or have a calcareous horizon at some depth or both. In the transition zone to the chernozemic region of North America, chernozemic braun erde (brown earth) and chernozemic lessivé (prairie) are common, and in the Appalachian Mountains, brun acide.

Chernozemic regions. A grassland vegetation with chernozemic soils occurs when spring is not dry, winter is relatively cold, and leaching rainfall is low. Calcareous, gypsic, or salic horizons are frequent. Solonchaks, solonetz, and planosols—more or less gleisolic and chernozemic—are common in poorly drained depressions.

The three chernozemic regions

The most important chernozemic regions of the world are the Danube Basin, southern Russia and Ukraine and part of Central Asia, the Great Plains of the United States and the prairies of Canada, and the Pampas region of Argentina. But there are some differences among them. The parent materials of Russian chernozems are more calcareous and the leaching rainfall is lower (for the same humidity index). Consequently, soils are richer in lime; effervescence with hydrogen chloride begins just below the humic horizon or within it; clay eluviation and planosolic chernozems are rare; and the humic horizon is, in general, deeper and richer in organic matter. In the North American Great Plains, parent materials are less calcareous, lime occurs at greater depth or not at all, clay eluviation is more frequent, planosolic chernozems more common, and the humic horizon is in general shallower and poorer in organic matter. In Argentina, parent materials are chiefly volcanic; cation-exchange capacity is unusually high (andic chernozemic); much sodium is released by weathering; clay eluviation is frequent; and planosolic chernozems, solonetz, and solonchaks are common.

The general pattern of soil distribution in chernozemic regions is chernozemic soils in well-drained areas, with planosols, solonetz, and solonchaks, more or less chernozemic and gleisolic, in areas of deficient drainage.

Cinnamonic regions. In climates with dry seasons where soil is dried thoroughly to considerable depth, fine crystallites or irreversibly dehydrated iron sesquioxides are formed, giving reddish colours and cinnamonic soils. In cinnamonic soils organic matter decays rapidly and does not accumulate on soil surface. Humus originates chiefly in roots and it is well distributed along the profile. Many cinnamonic regions have a Mediterranean climate. Those with a monsoon climate are usually transitional to brunisolic, chernozemic, desertic, or kaolinitic regions. Except for Australia, a Mediterranean climate is associated with mountainous topography and high frequency of limestones and volcanic materials; moreover, leaching rainfall varies considerably from zero to very high figures. Brown cinnamonic (bruns mediterraneans) and red cinnamonic (rouges mediterraneans) are common in areas with high leaching rainfall or ferruginous materials or both. They are sometimes acid and rich in 1:1 clays, being intergrades to kaolisols (krasnozems).

Desert regions. Vegetation is very scarce in the desert, and soil erosion is very severe from wind and occasional rains; weathering is slow and shallow, and leaching is almost absent. Soils are very poor in organic matter and do not have a humic horizon. Because sodium elimination is difficult, autochthonous soils are usually planosols and solonetz with very thin horizons (mini-planosols and mini-solonetz); a calcareous horizon usually underlies the B horizon. Soil is often covered by coarse material (desert pavement), which partly results from wind erosion and protects the soil. In such absolute deserts as the Chilean desert, however, products of weathering remain in the place of their formation and a loosely cemented desert crust is formed. Materials eroded from higher land accumulate in depressions, forming dunes and alluvial soils. Where waters accumulate, planosolic, solonetzic, and saline soils abound, and "salares" also are frequent. There are many good alluvial soils, however (rendzinas, chernozemic, etc.), and some of them are gleisolic.

Occurrence of mini-horizons

The general pattern of soil distribution may be summarized as follows: high plateaus and plains areas exhibit autochthonous soils, mini-planosols, mini-solonetz, and solonchaks with salares (salt accumulations); slopes exhibit lithosols; and depressions exhibit dunes and alluvial soils, with many solonetz, solonchak, serozem, and gleisolic soils with salares.

A distinction may be made between American (North and South), Asian, and African deserts. American deserts received much volcanic ash from active volcanoes, and mini-solonetz, mini-planosols, and duripans are plentiful. In Asia, calcareous materials abound, and raw (undifferentiated) soils that produce effervescence with hydrogen chloride are common. In Africa and Australia, palaeosols formed from kaolinitic materials are frequent.

Kaolinitic regions. The high leaching rainfall of tropical countries, even though there may be a long dry season, results in formation of 1:1 clays. Soils are usually acid in the lower horizons, whereas the upper horizon may be eutrophic. The difference between eutrophic and dystrophic soils, which is important from an agricultural point of view, seems well associated with the presence or absence of a dry season. A dry season, even a short one, activates decay of organic matter, favours fires, and interferes with the production of acid humus, which is the principal factor of acidification, several times more effective than water mixed with carbon dioxide.

Water-logging and iron concretions

In soils rich in 1:1 clays, waterlogging produces segregation of iron in the form of concretions or nodules. For that reason, laterite is common in waterlogged areas and, because it is impermeable, aggravates waterlogging and may reach the surface. In granite, gneiss, and similar materials, very common in Africa, the iron released by weathering may adhere to sand or gravel surfaces and cover them with a red coating. When soil is thoroughly dried periodically, such coatings become irreversibly dehydrated and protect the particle or gravel from further weathering, and a ferruginous horizon is formed. Such protection is insufficient in the upper horizon, where organic acids abound, and also in the lower horizons, which are waterlogged continuously; as a result, the ferruginous horizon is encountered at some depth. Some of the iron may come from the waterlogged horizons.

It is of interest that concretions of gravel and laterite contain unweathered material imprisoned in iron sesquioxides. Such material is virtually absent in the overlying and underlying horizons, indicating that both ferruginous horizons and laterite are formed at an early stage of weathering. This also indicates that stone lines of geologic origin may contribute to the formation of ferruginous horizons. It is, however, difficult to believe that there was a stone line parallel to soil surface each time a ferruginous horizon was formed. Ferruginous horizons and laterite are seldom found in soils formed from easily weatherable materials such as basalt, fine sand, and in continuously humid climates.

Young soils from volcanic ash are naturally ando in type. With time amorphous clays crystallize, cation-exchange capacity decreases, and soil becomes first terra roxa (intergrade ando-kaolinitic) and then kaolinitic. Terra roxa, however, is often made up of young soils formed from basalt and abounds in basaltic slopes, where erosion does not permit soil to become old (Misiones, Argentina; coffee region, Brazil).

Mountainous regions. Erosion is the main characteristic of mountainous regions. The soils are usually young, consisting of lithosols and rankers in slopes and alluvial soils in the valleys and surrounding the mountain plains.

Three kinds of mountainous regions may be distinguished: humid-temperate, dry, and humid-tropical. In humid-temperate climates organic matter accumulates and there is an abundance of brunisolic soils (braun erde, brun acide). There are also lessivé and podsols, but many podsols do not have an ashy horizon. The spodic horizon, rich in amorphous clays, is overlain by a gray-brown horizon (brown podsols); and limestones often form lithosolic rendzinas. Acid humus and peaty soils, more or less gleisolic, are common in depressions.

In dry but not desertic mountains, grassland vegetation, sometimes scattered with woody plants, is common. Lithosols and other raw soils and rendzinas, basically cinnamonic and chernozemic, are widespread, with saline, solonetzic, planosolic, and organic, relatively gleisolic soils in depressions.

In humid-tropical mountains the growing season is longer and the growth index much higher; soils are deeper and richer in organic matter (some giant podsols have been found); otherwise they resemble those of humid-temperate climate of the same summer type. At lower altitudes kaolinitic weathering begins and forms kaolinitic soils; erosion, however, results in soils remaining younger than in the lowlands, eutrophic soils and terra roxa are more frequent, and soils are also richer in organic matter. These characteristics partly account for the higher density of population in mountainous regions.

Intermediate and modified regions. Passage from one type of soil region to another is gradual, and intermediate regions can be recognized. Moreover, preponderance of allochthonous alluvial soils, waterlogging, or presence of a particular material justify separation of some regions from the main type.

Nature of transition belts

In the transition belt from podsolic to brunisolic regions, the prevailing soils are podsols and lessivé (gray-brown podsolic), and the distribution is chiefly determined by parent material. In the transition between podsolic and chernozemic regions, Russian authors recognize a forest-steppe zone with gray-wooded soils (lessivé with somewhat ashy horizon, better saturated with bases, and sometimes calcareous at some depth). Near this forest-steppe region are many sod-podsols (podsols with rather well-developed humic horizon), attributed to cropping. An analogous transitional region in Canada occurs between the podsolic and chernozemic region.

The prairie soils in the United States are characteristic of the transition between chernozemic and brunisolic regions and may be considered as chernozemic lessivé or chernozemic braun erde. They are also found in the Danube Basin and other parts of Europe, and they combine characteristics of both the chernozemic and brunisolic, or lessivé groups. The region of red-yellow podsolic soils of the southeastern United States may be considered as a transitional kaolinitic-brunisolic region with soils usually lessivé and very acid, usually low in cation-exchange capacity in comparison to their clay content. In other parts of the world, the transition between brunisolic and kaolinitic coincides with mountainous topography or special parent materials. The transition between brunisolic and cinnamonic regions in Europe is marked by cinnamonic-brunisolic intergrades (brun mediterraneans, southern braun erde, etc.); that between chernozemic and cinnamonic regions in the United States by reddish chernozems. The transition between cinnamonic and desertic regions in the United States is marked by reddish brown soils. Comparable soils are encountered in many parts of Asia, Africa, and Australia. Planosolic, solonetzic, and saline soils are common in these transitional regions.

In the transition between cinnamonic and kaolinitic regions, materials rich in easily weatherable minerals give soils with 2:1 clays, whereas 1:1 clays predominate in the opposite case, a condition that prevails in many parts of the Mediterranean Basin with high leaching rainfall and mild winters (*e.g.*, Portugal). In Africa and Australia, however, many kaolisols of such a transitional region are palaeokaolisols.

Volcanic regions. Volcanic ashes have such great influence on soil formation that volcanic regions constitute a special group. Ando soils are frequent in volcanic regions. A distinction may be made between humid volcanic regions with acid ando, often deficient in phosphorus, and dry volcanic regions with soils well provided with bases, often chernozemic. Planosolic, solonetzic, and saline soils, and duripans or calcareous horizons, are frequent in dry volcanic regions and still more so in desertic ones, where planosols and solonetz become mini-planosols and mini-solonetz. Ando soils are usually fertile, and as a result settlements grow around volcanoes in spite of dangers. The most important volcanic soil regions of the world are: the Cordillera (western mountain ranges) of North Amer-

Humid, and dry volcanic soils

ica; the entire Pacific coast of North and South America; the Great Basin of North America located within the Cordillera to the west of the Rocky Mountains; the Andes mountain ranges of South America, the plains to the north and west, and many lands to the east; all of Patagonia, the Pampas region of Argentina, and to a certain extent the Chaco; many parts of the Mediterranean Basin, especially in Italy; Indonesia, the Philippines, Japan, and Kamchatka; and some parts of Ethiopia, Kenya, Uganda, Rwanda, Burundi, and Cameroon.

Alluvial and waterlogged regions. Some regions have alluvial soils of materials that came from regions of different climate. The most conspicuous examples are the alluvial region of northern Italy, one of the better agricultural regions of the world, and those of northern India and Indochina, which nourish a dense population or export much food. In Australia, and to a certain extent in Africa, are regions in which the soils have been formed from kaolinitic materials developed under different conditions; these regions may be called paleokaolinitic.

In some regions waterlogging is so common that the soils may be called gleisolic. Some of them are podsolic; for example, the southwestern coast of Hudson Bay in Canada, most of Finland, and most of northern Russia. Others are chernozemic, like the "Pampa Deprimida" of Buenos Aires, Argentina. Still others are found in the transition region between chernozemic and kaolinitic soils of southern Brazil, Uruguay, and Corrientes, Argentina. Soils vary as a consequence: gleisolic humus podsols and peat in the first case; gleisolic chernozemic planosols and solonetz in the second; and chernozemic lessivé (prairie), chernozemic gleisolic (humic gley), dark clays (vertisols), planosols, acid gleisolic (low humic gley), and kaolinitic podsolic (red-yellow podsolic), with krasnozems and terra roxa on basaltic hills in the third.

Some tropical regions have relatively young soils, many of which have been formed from calcareous materials. This is the case in the Yucatán Peninsula of Mexico, in the West Indies, and on the coral islands of Oceania. Their soils are predominantly brunisolic-kaolinitic. (J.Pa.)

BIBLIOGRAPHY. **General texts.** Wide-ranging introductions include R.E. WHITE, *Introduction to the Principles and Practice of Soil Science,* 2nd ed. (1987); MILO I. HARPSTEAD, FRANCIS D. HOLE, and WILLIAM BENNETT, *Soil Science Simplified,* 2nd ed. (1988); DELVIN SEYMOUR FANNING and MARY CHRISTINE BALLUFF FANNING, *Soil: Morphology, Genesis, and Classification* (1989); SHEILA ROSS, *Soil Processes: A Systematic Approach* (1989); and NYLE C. BRADY, *The Nature and Properties of Soils,* 10th ed. (1990).

Soil formation. Processes of soil genesis are detailed in HANS JENNY, *The Soil Resource: Origin and Behavior* (1980); and S.W. BUOL, FRANCIS D. HOLE, and R.J. MCCRACKEN, *Soil Genesis and Classification,* 3rd ed. (1989).

Soil chemistry. D.J. GREENLAND and M.H.B. HAYES (eds.), *The Chemistry of Soil Processes* (1981), is a standard text. Also of interest on this topic are HINRICH L. BOHN, BRIAN L. MCNEAL, and GEORGE A. O'CONNOR, *Soil Chemistry,* 2nd ed. (1985); WILLARD L. LINDSAY, *Chemical Equilibria in Soils* (1979); GARRISON SPOSITO, *The Chemistry of Soils* (1989); F.J. STEVENSON, *Humus Chemistry: Genesis, Composition, Reactions* (1982); and J.B. DIXON and S.B. WEED (eds.), *Minerals in Soil Environments,* 2nd ed. (1989).

Nutrient cycles. The nutrient cycles in soil and the interactions between soils and plants are explained in J.L. HARLEY and R. SCOTT RUSSELL (eds.), *The Soil-Root Interface* (1979); STEPHEN T. TRUDGILL, *Soil and Vegetation Systems,* 2nd ed. (1988); and RAYMOND W. MILLER and ROY L. DONAHUE, *Soils: An Introduction to Soils and Plant Growth,* 6th ed. (1990). P.H. NYE and P.B. TINKER, *Solute Movement in the Soil-Root System* (1977), assumes some background knowledge.

Agriculture. Information on soil science as it pertains to agriculture is available in authoritative works by P.B. TINKER (ed.), *Soils and Agriculture* (1980); and EDWARD J. RUSSELL, *Russell's Soil Conditions and Plant Growth,* 11th ed. edited by ALAN WILD (1988). D.J. GREENLAND (ed.), *Characterization of Soils in Relation to Their Classification and Management for Crop Production* (1981), describes the productivity of the soils of certain parts of the humid tropics. R.J. HANCE (ed.), *Interactions Between Herbicides and the Soil* (1980), evaluates the relationship in terms of agricultural practice and the law.

Soil erosion and conservation. NORMAN HUDSON, *Soil Conservation,* 2nd ed. (1981), is a respected text. FREDERICK R. TROEH, J. ARTHUR HOBBS, and ROY L. DONAHUE, *Soil and Water Conservation for Productivity and Environmental Protection* (1980), provides an introduction for the nontechnical reader. Other useful texts are M.J. KIRKBY and R.P.C. MORGAN (eds.), *Soil Erosion* (1980); and S.A. EL-SWAIFY, W.C. MOLDENHAUER, and ANDREW LO (eds.), *Soil Erosion and Conservation* (1985). The following consider aspects of soil erosion: MARGARET R. BISWAS and ASIT K. BISWAS, *Desertification* (1980); SANDRA S. BATIE, *Soil Erosion: Crisis in America's Croplands?* (1983); and RATTAN LAL, *Soil Erosion in the Tropics: Principles and Management* (1990).

Life in the soil. Information on biotic components of soils is available in MARTIN ALEXANDER, *Introduction to Soil Microbiology,* 2nd ed. (1977); and B.N. RICHARDS, *Introduction to the Soil Ecosystem* (1974).

Geomorphology. K.S. RICHARDS, R.R. ARNETT, and S. ELLIS (eds.), *Geomorphology and Soils* (1985); and PETER W. BIRKELAND, *Pedology, Weathering, and Geomorphological Research* (1974), explain the connections between establishing soil profiles and interpreting geomorphic history. (Ed.)

The Solar System

The solar system consists of the Sun—an average star in the Milky Way Galaxy—and those bodies orbiting around it: 9 planets, more than 100 known planetary satellites (moons), countless asteroids and their fragments, comets, and vast reaches of gas and dust known as the interplanetary medium. This article surveys briefly the vast body of knowledge about the solar system and traces the progress in theories of its origin.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 131, 132, 133, 211, 212, 213, and 10/32, and the *Index*.

This article is divided into the following sections:

An overview of the solar system 452
- Major components and characteristics 452
- Origin of the solar system 454

The Sun 456
- Physical properties 456
- Internal structure 456
- Solar atmosphere 458
- Solar activity 460
- History of observation 463

The major planets and their satellites 464
- Mercury 464
  - Basic astronomical data
  - The atmosphere
  - The magnetic field and magnetosphere
  - Character of the surface
  - Origin and evolution
- Venus 469
  - Basic astronomical data
  - The atmosphere
  - Interaction with the solar wind
  - Character of the surface
  - Interior structure and geologic evolution
  - Observations from Earth
  - Spacecraft exploration
- Earth 477
  - Basic planetary data
  - The atmosphere and hydrosphere
  - The outer shell
  - The interior
  - The geomagnetic field and magnetosphere
- The Moon 482
  - Distinctive features
  - Principal characteristics of the Earth-Moon system
  - Motions of the Moon
  - The atmosphere
  - The lunar surface
  - The lunar interior
  - Origin and evolution
  - Lunar exploration
- Mars 493
  - Basic astronomical data
  - Early telescopic observations
  - General appearance
  - The atmosphere
  - The polar caps
  - Character of the surface
  - The interior
  - Meteorites from Mars
  - The satellites
  - Spacecraft exploration
  - The mapping of Mars
  - The question of life on Mars
- Jupiter 502
  - Basic astronomical data
  - The outer layers
  - The interior
  - The satellites and ring system
  - Theories of the origin of the Jovian system
- Saturn 508
  - Principal characteristics
  - The atmosphere
  - Interior structure and composition
  - Magnetic field and magnetosphere
  - The satellites and rings
  - History of observation
- Uranus 515
  - Principal characteristics
  - The atmosphere
  - Interior structure and composition
  - Magnetic field and magnetosphere
  - The satellites and rings
  - History of observation
- Neptune 520
  - Principal characteristics
  - The atmosphere
  - Interior structure and composition
  - Magnetic field and magnetosphere
  - The satellites and rings
  - History of observation
- Pluto 525
  - Basic astronomical data
  - The atmosphere
  - The surface and interior
  - Pluto's moon
  - Discoveries of Pluto and Charon
  - Origin of Pluto and Charon
  - Pluto's status as a planet

Other constituents of the solar system 528
- Asteroids 528
  - Historical survey of major asteroid discoveries
  - Orbits of asteroids
  - The nature of asteroids
  - The origin and evolution of the asteroids
- Comets 533
  - General considerations
  - Historical survey of comet observations and studies
  - Motion and discovery of comets
  - Cometary statistics
  - The nature of comets
  - Cometary models
  - Origin and evolution of comets
- Meteoroids, meteors, and meteorites 542
  - Meteoroids
  - Meteors
  - Meteorites

Bibliography 553

AN OVERVIEW OF THE SOLAR SYSTEM

Major components and characteristics

Located at the centre of the solar system, and affecting the motion of all the other bodies through its gravitational force, is the Sun, which in itself contains more than 99 percent of the mass of the system. The planets, in order of distance outward from the Sun, are Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and Pluto. Four planets—Jupiter through Neptune—have ring systems, and all but Mercury and Venus have one or more moons.

Any natural solar system object other than a planet or moon is called a small body; these include asteroids, meteoroids, and comets. Most of the several thousand asteroids, or minor planets, orbit between Mars and Jupiter in a flat ring called the asteroid belt. Asteroids can have moons of their own, as was revealed by spacecraft and telescopic observations in the 1990s. The myriad fragments of asteroids and other small pieces of solid matter (smaller than a few tens of metres across) that populate interplanetary space are often termed meteoroids to distinguish them from the larger asteroidal bodies.

The solar system's several billion comets are found main-

ly in two distinct reservoirs. The more distant Oort cloud is a spherical shell surrounding the solar system at a distance of approximately 50,000 astronomical units (AU)—more than 1,000 times the distance of Pluto's orbit. The nearer reservoir, the Kuiper belt, is a disk-shaped zone extending 30–50 AU from the Sun, beyond the orbit of Neptune but including that of Pluto. (One astronomical unit is the average Earth-Sun distance—about 150 million kilometres.) Although it is traditionally called a planet and is likely to remain classified as such, Pluto and its moon, Charon, are actually the largest members of the Kuiper belt. Pluto is a giant surviving representative of the icy solid bodies called planetesimals that accreted to form the cores of Neptune and Uranus. The Centaurs, a population of comet nuclei having diameters as large as 200 kilometres, orbit the Sun between Jupiter and Neptune, probably having been gravitationally perturbed inward from the Kuiper belt. The interplanetary medium—an exceedingly tenuous plasma (ionized gas) laced with dust particles—extends outward from the Sun to great distances.

Status of Pluto as a planet

ORBITS

All the planets, asteroids, and icy bodies in the Kuiper belt move around the Sun in elliptical orbits in the same direction that the Sun rotates. This motion is termed prograde, or direct, motion. The icy bodies in the Oort cloud are in orbits having random directions, corresponding to their spherical distribution around the plane of the planets.

The shape of an object's orbit is defined in terms of its eccentricity. For a perfectly circular orbit, the eccentricity is 0; with increasing elongation of the orbit's shape, the eccentricity increases toward a value of 1. Of the planets, Venus and Neptune have the most circular orbits, with eccentricities of 0.007 and 0.009, respectively. Mercury and Pluto, the closest and most distant planets, have the highest eccentricities, at 0.21 and 0.25, respectively. Another defining attribute of an object's orbit around the Sun is its inclination, which is the angle that it makes with the plane of Earth's orbit—the ecliptic plane. Again, of the planets, Mercury and Pluto have the greatest inclinations: Mercury's orbit is inclined at 7° and Pluto's at 17°. The orbits of the small bodies generally have both higher eccentricities and higher inclinations than those of the planets.

PLANETS AND THEIR MOONS

The planets can be divided into two distinct categories on the basis of their densities. The inner, or terrestrial, planets—Mercury, Venus, Earth, and Mars—have rocky compositions and densities greater than three grams per cubic centimetre. (Water has a density of one gram per cubic centimetre.) In contrast, the outer planets, also called the Jovian, or giant, planets—Jupiter, Saturn, Uranus, and Neptune—are large objects with densities less than two grams per cubic centimetre; they are composed mostly of hydrogen and helium (Jupiter and Saturn) or of ice, rock, hydrogen, and helium (Uranus and Neptune). Pluto is unique—an icy, low-density body smaller than Earth's Moon, more similar to comets or to the large icy moons of the outer planets than to any of the planets themselves. (For comparisons of the distance, density, and composition of the planets and selected moons, see Table 1.)

Table 1: Compositional Data for Selected Objects

object	distance from Sun (AU)*	density (g/cm³)	composition
Mercury	0.4	5.4	iron, nickel, silicates
Venus	0.7	5.2	silicates, iron, nickel
Earth	1.0	5.5	silicates, iron, nickel
Moon	1.0	3.3	silicates
Mars	1.5	3.9	silicates, iron, sulfur
Jupiter	5.2	1.3	hydrogen, helium
Io	5.2	3.6	silicates
Europa	5.2	3.0	silicates, water, ice crust
Ganymede	5.2	1.9	water ice, silicates
Callisto	5.2	1.8	water ice, silicates
Saturn	9.5	0.7	hydrogen, helium
Titan	9.5	1.9	water ice, silicates
Uranus	19.2	1.3	ices, silicates, hydrogen, helium
Neptune	30.1	1.7	ices, silicates, hydrogen, helium
Triton	30.1	2.0	water ice, silicates, organics
Pluto	39.5	2.0	water ice, silicates, organics

*One astronomical unit (AU) is defined as the mean distance of Earth from the Sun and is equal to 149.6 million kilometres.

The relatively small inner planets have solid surfaces, lack rings, and have few or no moons. The atmospheres of Venus, Earth, and Mars are highly oxidized; on Venus and Mars, carbon dioxide is the dominant gas, while Earth's atmosphere is 21 percent molecular oxygen. Of the inner planets, only Earth has a strong magnetic field, which shields it from the interplanetary medium.

The outer planets are much more massive than the terrestrial planets and have immense atmospheres composed mainly of hydrogen and helium. They have no solid surfaces, however, and their densities are so low that one of them, Saturn, would actually float in water. Each of the outer planets has a magnetic field, a ring system, and many known moons. Pluto is again the exception, having no rings and only a single moon.

ASTEROIDS AND COMETS

The asteroids and comets are remnants of the planet-building process in the inner and outer solar system, respectively. The asteroid belt is home to rocky bodies ranging in size from the largest known asteroid, Ceres, with a diameter of roughly 920 kilometres, to microscopic dust particles that are dispersed throughout the belt. Some asteroids travel in paths that cross the orbit of Earth, providing opportunities for collisions with the planet. The rare collisions of relatively large objects (those with a radius of more than 10 kilometres) with Earth can be devastating, as in the case of the asteroid impact that is thought to have been responsible for the massive extinction of species at the end of the Cretaceous Period 65 million years ago (see DINOSAURS: *The search for dinosaurs: Extinction*). More commonly, the impacting objects are much smaller, reaching Earth's surface as meteorites. Observations of asteroids from Earth, which have been confirmed by spacecraft flybys, indicate that some are mainly metal (principally iron), others are stony, and still others are rich in organic compounds, resembling the carbonaceous chondrite meteorites.

Collisions of asteroids with Earth

The physical characteristics of comet nuclei are fundamentally different from those of asteroids. Ices are their main constituent, mainly in the form of frozen water, but frozen carbon dioxide, carbon monoxide, and other ices are also present. These cosmic ice balls are laced with rock dust and a rich variety of organic compounds. Typically they are irregularly shaped and a few kilometres across.

As comet nuclei trace out the parts of their orbits closest to the Sun, they are warmed through solar heating and begin to shed gases and dust, which form the familiar fuzzy-looking comas and long, wispy tails. The gas dissipates into space, but the grains of silicates and organic compounds remain to orbit the Sun along paths close to that of the parent comet. When Earth's path around the Sun intersects one of these dust-populated orbits, a meteor shower occurs. During such an event, nighttime observers may see tens to hundreds of so-called shooting stars per hour as the dust grains burn up in the upper atmosphere of Earth.

THE INTERPLANETARY MEDIUM

In addition to particles of debris, the space through which the planets travel contains protons, electrons, and ions of the abundant elements, all streaming outward from the Sun in the form of the solar wind. Occasional giant solar flares, short-lived eruptions on the Sun's surface, expel matter (along with high-energy radiation) that contributes to this interplanetary medium.

At the start of the 21st century, astronomers had yet to determine exactly where the boundary between the interplanetary medium and the interstellar medium lies. This boundary is called the heliopause. Four spacecraft, Pioneers 10 and 11 and Voyagers 1 and 2, already have passed the orbit of Pluto with velocities high enough to allow them to escape the solar system. The chances appear good that at least the two Voyagers will remain operational long enough to cross the heliopause and return measurements of the properties of interstellar space.

Search for the solar system's edge

Origin of the solar system

As the amount of data on the planets, moons, comets, and asteroids has grown (for the orbital and physical characteristics of these objects, see the sections following), so too have the problems faced by astronomers in forming theories of the origin of the solar system. In the ancient world, theories of the origin of Earth and the objects seen in the sky were certainly much less constrained by fact. Indeed, a scientific approach to the origin of the solar system became possible only after the publication of Isaac Newton's laws of motion and gravitation in 1687. Even then, many years elapsed while scientists struggled with applications of Newton's laws to explain the apparent motions of planets, moons, comets, and asteroids. Meanwhile, the first semblance of a modern theory was proposed by the German philosopher Immanuel Kant in 1755.

EARLY THEORIES

The Kant-Laplace hypothesis. Kant's central idea was that the solar system began as a cloud of dispersed particles. He assumed that the mutual gravitational attractions of the particles caused them to start moving and colliding, at which point chemical forces kept them bonded together. As some of these aggregates became larger than others, they grew still more rapidly, ultimately forming the planets. Kant's model, however, does not account for planets moving around the Sun in the same direction and in the same plane, as they are observed to do, nor does it explain the revolution of planetary satellites.

A significant step forward was made by the mathematician Pierre-Simon Laplace of France some 40 years later. Besides publishing a monumental treatise on celestial mechanics, Laplace wrote a popular book on astronomy, with an appendix in which he made some suggestions about the origin of the solar system. It is for this relatively minor work that he is best remembered.

Laplace's model begins with the Sun already formed and rotating and its atmosphere extending beyond the distance at which the farthest planet would be created. Knowing nothing about the source of energy in stars, Laplace assumed that the Sun would start to cool as it radiated away its heat. In response to this cooling, as the pressure exerted by its gases declined, the Sun would contract. Owing to the law of conservation of angular momentum, the decrease in size would be accompanied by an increase in the Sun's rotational velocity. Centrifugal acceleration would push the material in the atmosphere outward, while gravitational attraction would pull it toward the central mass; when these forces just balanced, a ring of material would be left behind in the plane of the Sun's equator. This process would have continued through the formation of several concentric rings, each of which then would have coalesced to form a planet. Similarly, a planet's moons would have originated from rings produced by the forming planets.

Strengths of Laplace's model

Laplace's model led naturally to the observed result of planets revolving around the Sun in the same plane and in the same direction as the Sun rotates. Because Laplace's theory incorporated Kant's idea of planets coalescing from dispersed material, their approaches are often combined in a single model called the Kant-Laplace nebular hypothesis. This model for solar system formation was widely accepted for about 100 years. During this period, the apparent regularity of motions in the solar system was contradicted by the discovery of asteroids with highly eccentric orbits and moons with retrograde orbits. Another problem was the fact that, whereas the Sun contains 99.9 percent of the mass of the solar system, the planets (principally the outer planets) carry more than 99 percent of the angular momentum. For the solar system to conform to this theory, either the Sun should be rotating more rapidly or the planets should be revolving around it more slowly.

20th-century developments. In the early decades of the 20th century, several scientists decided that the deficiencies of the nebular hypothesis made it no longer tenable. Independently they developed variations on the idea that the planets were formed catastrophically—*i.e.*, by a close encounter of the Sun with another star. The basis of this model was that, as the two bodies passed, material was drawn from one or both stars, and this material later coalesced to form planets. A discouraging aspect of the theory was the implication that the formation of solar systems in the Galaxy must be extremely rare, because sufficiently close encounters between stars would occur very seldom.

Fatal flaw in catastrophe theories

The next significant development took place in the mid-20th century, as scientists acquired a more mature understanding of the processes by which stars themselves must form and of the behaviour of gases within and around stars. They realized that hot gaseous material stripped from a stellar atmosphere would simply dissipate in space. Hence, the basic idea that a solar system could form through stellar encounters was untenable. Furthermore, the growth in knowledge about the interstellar medium—the gas and dust distributed in the space separating the stars—indicated that large clouds of such matter exist and that stars form there. Planets must somehow be created in the same process that forms the stars. This awareness encouraged scientists to reconsider certain basic processes that resembled the earlier notions of Kant and Laplace.

MODERN IDEAS

The current approach to the origin of the solar system treats it as part of the general process of star formation. As observational information has steadily increased, the field of plausible models for this process has narrowed.

Formation of the solar nebula. The favoured paradigm for the origin of the solar system begins with the gravitational collapse of part of an interstellar cloud of gas and dust, with an initial mass only 10–20 percent greater than the present mass of the Sun. Because the cloud is revolving around the centre of the Galaxy, the parts more distant from the centre are moving more slowly than the nearer parts. Hence, as the cloud collapses, it starts to rotate, and to conserve angular momentum, its speed of rotation increases as it continues to contract. The cloud also flattens, because it is easier for matter to follow the attraction of gravity perpendicular to the plane of rotation than along it, where the opposing centrifugal force is greatest. The result at this stage, as in Laplace's model, is a disk of material formed around a central condensation.

This configuration, commonly referred to as the solar nebula, resembles the shape of a typical spiral galaxy on a much reduced scale. As gas and dust are pulled in toward the central condensation, their potential energy is converted to kinetic energy, and the temperature of the material rises. Ultimately the material becomes hot enough for nuclear reactions to begin, thereby giving birth to a star.

Meanwhile, the material in the disk collides, coalesces, and gradually forms larger and larger objects, as in Kant's theory. Because most of the grains of material have nearly identical orbits, their collisions are relatively mild, which allows them to stick and remain together. Thus larger agglomerations of particles are gradually built up.

Differentiation into inner and outer planets. At this stage the individual accreting objects in the disk show differences in growth and composition that depend on their distance from the hot central mass. Close to the nascent Sun, temperatures are too high for water to condense from a gas to ice, but at distances of present-day Jupiter and beyond, water ice can form. Because the water molecule is the second most abundant molecule in the universe (after molecular hydrogen), objects forming at these lower temperatures can acquire much more mass in the form of solid material than objects forming closer to the Sun. Once such an accreting body achieves about 10 times the present mass of Earth, its gravity can attract and retain large amounts of even the lightest elements, hydrogen and helium, from the solar nebula. These are the two most abundant elements in the universe, and so planets forming in this region can become very massive indeed.

Key role of water in the development of giant planets

This simple picture can explain the extensive differences observed between the inner and outer planets. The inner planets formed at temperatures that were too high to allow the abundant volatile substances—those with comparatively low freezing temperatures—like water, carbon dioxide, and ammonia to condense to their ices. They therefore remained small rocky bodies. In contrast, the low-density, gas-rich outer planets formed at distances beyond the min-

imum distance from the Sun at which water ice could have condensed, about 150 K (−190° F, −120° C). (See Table 1.) The effect of the temperature gradient in the solar nebula can be seen today in the increasing fraction of condensed volatiles in solid bodies as their distance from the Sun increases. As the nebular gas cooled, the first solid materials to condense were grains of metal-containing silicates, the basis of rocks. This was followed, at larger distances from the Sun, by formation of the ices. In the inner solar system, Earth's Moon, with a density of 3.3 grams per cubic centimetre, is a satellite composed of silicate minerals. In the outer solar system are low-density moons such as Saturn's Tethys. With a density of about 1 gram per cubic centimetre, this object must consist mainly of water ice. At still farther distances, satellite densities rise again, but only slightly, presumably because they incorporate denser solids, such as carbon dioxide, that freeze at even lower temperatures.

Despite its apparent logic, this scenario has received strong challenges since the early 1990s. One in particular has come from the discovery of other solar systems, many of which contain giant planets orbiting very close to their stars in seeming defiance of the minimum radius for ice condensation. (See *Studies of other solar systems,* below.)

Although a number of problems remain to be resolved, the solar nebula model of Kant and Laplace appears basically correct. Support comes from observations at infrared and radio wavelengths, which have revealed disks of matter around young stars. These observations also suggest that planets form in a remarkably short time. The collapse of an interstellar cloud into a disk should take about one million years. The thickness of this disk is determined by the gas it contains, as the solid particles that are forming rapidly settle to the disk's midplane, in times ranging from 100,000 years for 1-micrometre (0.00004-inch) particles to just 10 years for 1-centimetre (0.4-inch) particles. As the local density increases at the midplane, the opportunity becomes greater for the growth of particles by collision. As the particles grow, the resulting increase in their gravitational fields accelerates further growth. Calculations show that objects 10 kilometres in size will form in just 1,000 years. Such objects are large enough to be called planetesimals, the building blocks of planets.

Rapidity of planet formation

Later stages of planetary accretion. Continued growth by accretion leads to larger and larger objects. The energy released during accretionary impacts would be sufficient to cause vaporization and extensive melting, transforming the original primitive material that had been produced by direct condensation in the nebula. Theoretical studies of this phase of the planet-forming process suggest that several bodies the size of the Moon or Mars must have formed in addition to the planets found today. Collisions of these giant planetesimals with the planets would have had dramatic effects and could have produced some of the anomalies seen today in the solar system—for example, the extremely slow and retrograde rotation of Venus and the strangely high density of Mercury. A collision of a planetesimal with Earth also could have formed the Moon.

Studies of isotopes formed from the decay of radioactive elements with short half-lives, in both lunar samples and meteorites, have demonstrated that the formation of the inner planets, including Earth, and the Moon was essentially complete within 50 million years after the interstellar cloud collapsed. The bombardment of planetary and satellite surfaces by leftover debris continued intensively for another 600 million years, but these impacts contributed only a few percent of the mass of any given object.

Formation of the outer planets and their moons. This general scheme of planet formation—the building up of larger masses by the accretion of smaller ones—occurred in the outer solar system as well. There, the planets could grow so large that their composition came to resemble that of the Sun itself. Each planet started with its own "subnebula," forming a disk with a central condensation. The so-called regular satellites of the outer planets, which today have nearly circular orbits close to the equatorial planes of their respective planets and orbital motion in the same direction as the planet's rotation, formed from this disk. The irregular satellites—those having orbits with high inclination, high eccentricity, or both and sometimes also retrograde motion—must represent objects formerly in orbit around the Sun that were gravitationally captured by their respective planets.

It is interesting that the density distribution of Jupiter's Galilean satellites, its four largest regular moons, mirrors that of the planets in the solar system at large. The two Galilean moons closest to the planet, Io and Europa, are rocky bodies, while the more distant Ganymede and Callisto are half ice. Models for the formation of Jupiter suggest that this giant planet was sufficiently hot during its early history that ice could not condense in the circumplanetary nebula at the present position of Io. (See *Jupiter: Theories of the origin of the Jovian system,* below.)

The Galilean moons as a mirror of the solar system

The small bodies. At some point after most of the matter in the solar nebula had formed discrete objects, a sudden increase in the intensity of the solar wind apparently cleared the remaining gas and dust out of the system. Astronomers have found evidence of such strong outflows around young stars. The larger debris from the nebula remained, some of which is seen today as asteroids and comets. The rapid growth of Jupiter apparently prevented the formation of a planet between Jupiter and Mars; within this area remain the thousands of objects that make up the asteroid belt. The meteorites that are recovered on Earth, the great majority of which come from these asteroids, provide important clues to the conditions and processes in the early solar nebula.

The icy comet nuclei are representative of the planetesimals that formed in the outer solar system. Most are extremely small, but the Centaur object called Chiron—originally classified as a distant asteroid but now known to show characteristics of a comet—has a diameter estimated to be about 200 kilometres. Other bodies of this size have been observed in the Kuiper belt and may also reside in the Oort cloud. The objects occupying the Kuiper belt apparently formed in place, but calculations show that billions of icy planetesimals were gravitationally expelled by the giant planets from their vicinity as the planets formed. These objects became the population of the Oort cloud.

Formation of ring systems. The process of planetary ring formation remains a subject of intense research, although their existence is readily explainable. Each planet has a critical distance from its centre known as its Roche limit after Édouard Roche, the 19th-century French mathematician who first explained this concept. The ring systems of Jupiter, Saturn, Uranus, and Neptune lie inside the Roche limits of their respective planets. Within this distance the gravitational attraction of two small bodies for each other is smaller than the difference in the attraction of the planet for each of them. Hence, the two cannot accrete to form a larger object. Moreover, because a planet's gravitational field acts to disperse the distribution of small particles in a surrounding disk, the random motions that would lead to accretion by collision are minimized.

The Roche limit

The problem challenging astronomers is in understanding how and when the material making up a planet's rings reached its present position within the Roche limit and how the rings are radially confined. These processes are likely to be very different for the different ring systems.

Solution to the angular momentum puzzle. The angular momentum problem that defeated Kant and Laplace—why the planets have most of the solar system's angular momentum while the Sun has most of the mass—can now be approached in a cosmic context. All stars having masses that range from slightly above the mass of the Sun to the smallest known masses rotate more slowly than an extrapolation based on the rotation rate of stars of higher mass would predict. Accordingly, these sunlike stars show the same deficit in angular momentum as is found in the Sun itself.

The answer to how this loss could have occurred seems to lie in the solar wind. The Sun and other stars of comparable mass have outer atmospheres that are slowly but steadily expanding into space. Stars of higher mass do not exhibit such stellar winds. The loss of angular momentum associated with this loss of mass to space is sufficient to reduce the rate of the Sun's rotation. Thus the planets preserve the angular momentum that was in the original solar nebula,

but the Sun has gradually slowed down in the 4.6 billion years since it formed.

STUDIES OF OTHER SOLAR SYSTEMS

Astronomers have long wondered if the process of planetary formation has accompanied the birth of stars other than the Sun. The discovery of extrasolar planets—planets circling other stars—would help clarify their ideas of the formation of Earth's solar system by allowing them to study more than one example. Extrasolar planets have not been seen directly with Earth-based telescopes, but they can be observed indirectly by noting the gravitational effects that they exert on the motion of their parent stars—for example, slight wobbles produced in the parent star's motion through space or, alternately, small periodic changes in some property of the star's radiation, caused by the planet's tugging the star first toward and then away from the direction of Earth.

After decades of searching for extrasolar planets, astronomers in the early 1990s made the first confirmed discoveries, and within the next few years the count of stars known to have one or more orbiting planets exceeded 100. Included among these discoveries are systems comprising giant planets the size of several Jupiters orbiting their stars at distances closer than that of the planet Mercury to the Sun. Totally different from the Earth's solar system, they appear to violate a basic tenet of the formation process discussed above—that giant planets must form far enough from the hot central condensation to allow ice to condense. One solution has been to postulate that giant planets can form quickly enough to leave plenty of matter in the disk-shaped nebula between them and their stars. Tidal interaction of the planet with this matter can cause the planet to spiral slowly inward, stopping where disk material no longer is present because the star has consumed it.

Perplexing orbits of extrasolar giant planets

Most of the extrasolar planets discovered to date have masses similar to or greater than that of Jupiter. As techniques are developed for detecting smaller planets, astronomers will better understand how planetary systems, including the Sun's, form and evolve. (T.C.O.)

THE SUN

The Sun is the mother star around which the Earth revolves once a year. It is the source of heat, light, and life itself on the Earth. The Sun is classified as a G2 V star, with G2 standing for the second hottest stars of the yellow G class—of surface temperature about 6,000 kelvins (K)—and the V representing a main sequence, or dwarf, star, the typical star for this temperature class. (G stars are so called because of the prominence of a band of atomic and molecular spectral lines that the German physicist Joseph von Fraunhofer designated G.) The Sun exists in the outer part of the Milky Way Galaxy and was formed from material that had been processed inside a supernova. The Sun is not, as is often said, a small star. Although it falls midway between the biggest and smallest stars of its type, there are so many dwarf stars that the Sun falls in the top 5 percent of stars in the neighbourhood that immediately surrounds it.

Classification as a star

Physical properties

The radius of the Sun, $R_\odot$, is 109 times that of the Earth, but its distance from Earth is 215 $R_\odot$, so it subtends an angle of only ½° in the sky, roughly the same as that of the Moon. By comparison, the next closest star to the Earth is 250,000 times farther away, and its relative apparent brightness is reduced by the square of that ratio, or 62 billion times. The temperature of the Sun's surface is so high that no solid or liquid can exist there; the constituent materials are predominantly gaseous atoms, with a very small number of molecules. As a result, there is no fixed surface. The surface viewed from Earth, called the photosphere, is the layer from which most of the radiation reaches us; the radiation from below is absorbed and reradiated, and the emission from overlying layers drops sharply, by about a factor of six every 200 kilometres (124 miles). The Sun is so far from the Earth that this slightly fuzzy surface cannot be resolved, and so the limb (the visible edge) appears sharp.

The mass of the Sun, $M_\odot$, is 743 times the total mass of all the planets in the solar system and 330,000 times that of the Earth. All the interesting planetary and interplanetary gravitational phenomena are negligible effects in comparison to the force exerted by the Sun. Under the force of gravity, the great mass of the Sun presses inward, and to keep the star from collapsing, the central pressure outward must be great enough to support its weight. The density at the Sun's core is about 100 times that of water (roughly six times that at the centre of the Earth), but the temperature is at least 15,000,000 K, so the central pressure is at least 10,000 times greater than that at the centre of the Earth, which is 3,500 kilobars. The nuclei of atoms are completely stripped of their electrons, and at this high temperature they collide to produce the nuclear reactions that are responsible for generating the energy vital to life on Earth.

While the temperature of the Sun drops from 15,000,000 K at the centre to 6,000 K at the photosphere, a surprising reversal occurs above that point; the temperature drops to a minimum of 4,000 K, then begins to rise in the chromosphere, a layer about 7,000 kilometres high at a temperature of 8,000 K. During a total eclipse the chromosphere appears as a pink ring. The photosphere and chromosphere are shown in Figure 1. Above the chromosphere is a dim, extended halo called the corona, which has a temperature of 1,000,000 K and reaches far past the planets. Beyond a distance of $5R_\odot$ from the Sun, the corona flows outward at a speed (near the Earth) of 400 kilometres per second (km/s); this flow of charged particles is called the solar wind.

The Sun is a very stable source of energy; its radiative output, called the solar constant, is 137 ergs per square metre per second (ergs/m^2/sec), or 1.98 calories per square centimetre per minute (cal/cm^2/min), at the Earth and varies by no more than 0.1 percent. Superposed on this stable star, however, is an interesting 11-year cycle of magnetic activity manifested by regions of transient strong magnetic fields called sunspots.

Solar constant

Internal structure

ENERGY GENERATION AND TRANSPORT

The energy radiated by the Sun is produced during the conversion of hydrogen (H) atoms to helium (He). The Sun is at least 90 percent hydrogen by number of atoms, so the fuel is readily available. Since one hydrogen atom weighs 1.0078 atomic mass units, and a single helium atom weighs 4.0026, the conversion of four hydrogen atoms to one helium atom yields 0.0294 mass unit, which are all converted to energy, 6.8 million-electron volts (MeV), in the form of gamma (γ) rays or the kinetic energy of the products. If all the hydrogen is converted, 0.7 percent of the mass becomes energy, according to the Einstein formula $E = mc^2$, in which E represents the energy, m is the mass, and c is the speed of light. A calculation of the time required to convert all the hydrogen in the Sun provides an estimate of the length of time for which the Sun can continue to radiate energy. Converting 0.7 percent of the 2×10^{33} grams of hydrogen into energy that is radiated at 4×10^{33} ergs per second permits the Sun to shine for 3×10^{18} seconds, or 100 billion years at the present rate.

The process of energy generation results from the enormous pressure and density at the centre of the Sun, which makes it possible for nuclei to overcome electrostatic repulsion. (Nuclei are positive and thus repel each other.) Once in some billions of years a given proton (^{1}H, in which the superscript represents the mass of the isotope) is close enough to another to undergo a process called inverse beta-decay, in which one proton becomes a neutron and combines with the second to form a deuteron (^{2}D). This is shown symbolically on the first line of equation 1,

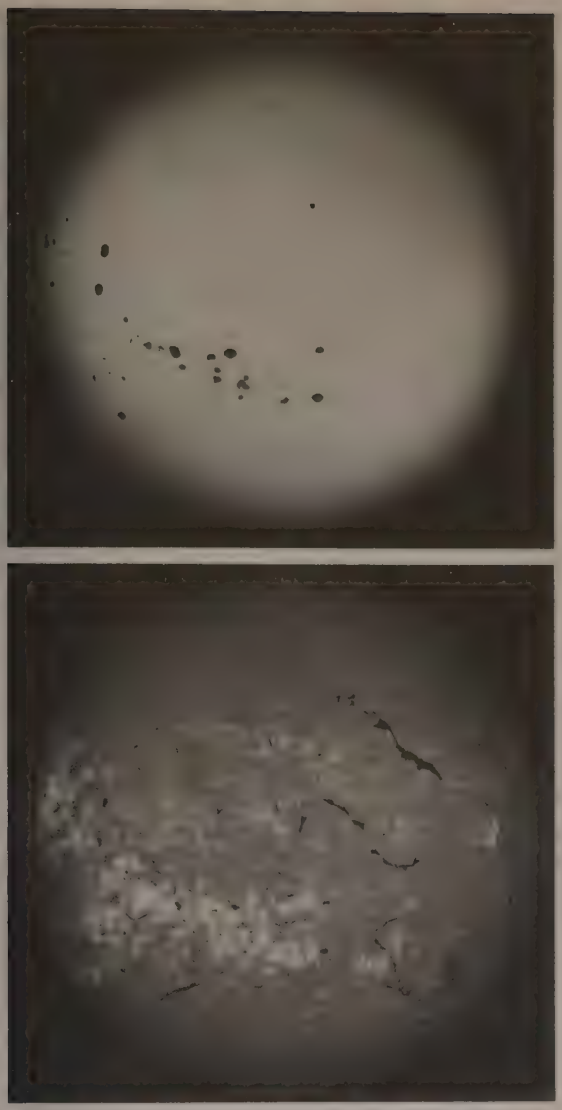


Figure 1: Image of the Sun in continuous light (top), which shows the photosphere, and another in the red Hα line of hydrogen at 6562.8 angstroms (bottom), which shows the chromosphere. The dark spots are sunspots, which in the Hα line are surrounded by bright patches called plages. The elongated dark clouds are filaments, lying high above the surface. At the limb these appear bright against the sky, where they are called prominences.
Big Bear Solar Observatory, California Institute of Technology

in which e^- is an electron and ν is a subatomic particle known as a neutrino.

$$\begin{aligned} e^- + {}^1\mathrm{H} + {}^1\mathrm{H} &\rightarrow {}^2\mathrm{D} + \nu + 1.44\ \mathrm{MeV} \\ {}^2\mathrm{D} + {}^1\mathrm{H} &\rightarrow {}^3\mathrm{He} + \gamma + 5.49\ \mathrm{MeV} \\ {}^3\mathrm{He} + {}^3\mathrm{He} &\rightarrow {}^4\mathrm{He} + {}^1\mathrm{H} + {}^1\mathrm{H} + 12.85\ \mathrm{MeV} \end{aligned} \quad (1)$$

While this is a rare event, hydrogen atoms are so numerous that it is the main solar energy source. Subsequent encounters (listed on the second and third lines) proceed much faster: the deuteron encounters one of the ubiquitous protons to produce helium-3 (^{3}He), and these in turn form helium-4 (^{4}He). The net result is that four hydrogen atoms are fused into one helium atom. The energy is carried off by gamma-ray photons (γ) and neutrinos, ν. Because the nuclei must have enough energy to overcome the electrostatic barrier, the rate of energy production varies as the fourth power of the temperature.

Equation 1 shows that for every two hydrogen atoms converted, one neutrino of average energy 0.26 MeV carrying 1.3 percent of the total energy released is produced. This produces a flux of 8×10^{10} neutrinos per square centimetre per second at the Earth. These neutrinos have an energy (less than 0.42 MeV) that is too low to be detected by present experiments, so there is considerable effort today to develop experiments that can detect them. Subsequent processes produce higher-energy neutrinos that have been detected in an experiment designed by the American scientist Raymond Davis and carried out deep in the Homestake gold mine in Lead, S.D., U.S. The number of these higher-energy neutrinos observed has been far smaller than would be expected from the known energy-generation rate, but experiments have established that these neutrinos do in fact come from the Sun. One possible reason for the small number detected is that the presumed rates of the subordinate process are not correct. Another more intriguing possibility is that the neutrinos produced in the core of the Sun interact with the vast solar mass and change to a different, undetectable kind of neutrino that cannot be observed. The existence of such a process would have great significance for nuclear theory, for it requires a small mass for the neutrino, which thus far is believed to be massless. Preliminary results from the new experiments show a flux of low-energy neutrinos within 30 percent of the expected values, so this appears to actually occur.

Neutrino detection

In addition to being carried away as neutrinos, which simply disappear into the cosmos, the energy produced in the core of the Sun takes two other forms as well. Some is released as the kinetic energy of product particles, which heats the gases in the core, while some travels outward as gamma-ray photons until they are absorbed and reradiated by the local atoms. Because the nuclei at the core are completely ionized, or stripped of their electrons, the photons are simply scattered there into a different path. The density is so high that the photons travel only a few millimetres before they are scattered. Farther out the nuclei have electrons attached, so they can absorb and reemit the photons, but the effect is the same: the photons take a so-called random walk outward until they escape from the Sun. The distance covered in a random walk is the average distance traveled between collisions (known as the mean free path) multiplied by the square root of the number of steps, in which a step is an interval between successive collisions. As the average mean free path in the Sun is about 10 centimetres (4 inches), the photon must take 5×10^{19} steps to travel 7×10^{10} centimetres. Even at the speed of light this process takes 10 million years, and so the light seen today was generated long ago. The final step from the Sun's surface to Earth, however, takes only eight minutes.

As photons are absorbed by the outer portion of the Sun, the temperature gradient increases and convection occurs. Great currents of hot plasma, or ionized gas, carry heat upward. These mass motions of conducting plasma in the convective zone, which constitutes approximately the outer 30 percent of the Sun, may be responsible for the sunspot cycle. The ionization of hydrogen plays an important role in the transport of energy through the Sun. Atoms are ionized at the bottom of the convective zone and are carried upward to cooler regions, where they recombine and liberate the energy of ionization. Just below the surface, radiation transport again becomes efficient, but the effects of convection are clearly visible in the photosphere.

Convective zone

EVOLUTION

The geologic record of the Earth and Moon reveals that the Sun has been shining at least four billion years. Considerable hydrogen has been converted to helium in the core, where the burning is most rapid. The helium remains there, where it absorbs radiation more readily than hydrogen. This raises the central temperature and increases the brightness. Model calculations conclude that the Sun becomes 10 percent brighter every billion years; hence it must now be at least 40 percent brighter than at the time of planet formation. This would produce an increase in the Earth's temperature, but no such effect appears in the fossil record. There may be compensating thermostatic effects in the atmosphere of the Earth, such as the greenhouse effect and cloudiness. The increase in solar brightness can be expected to continue as the hydrogen in the core is depleted and the region of nuclear

burning moves outward. At least as important for the future of the Earth is the fact that tidal friction will slow down the Earth's rotation until, in four billion years, its rotation will match that of the Moon, turning once in 30 of our present days.

The evolution of the Sun should continue on the same path as that taken by most stars. As the core hydrogen is used up, the nuclear burning will take place in a growing shell surrounding the exhausted core. The star will continue to grow brighter, and when the burning approaches the surface, the Sun will enter the red giant phase, producing an enormous shell that may extend as far as Venus or even the Earth. Fortunately, unlike more massive stars that have already reached this state, billions of years will pass before this catastrophe occurs.

Red giant phase

HELIOSEISMOLOGY

The structure of a star is uniquely determined by its mass and chemical composition. Unique models are constructed by varying the assumed composition with the known mass until the observed radius, luminosity, and surface temperature are matched. The process also requires assumptions about the convective zone. Such models can now be tested by the new science known as helioseismology.

Helioseismology is analogous to geoseismology: frequencies and wavelengths of various waves at the Sun's surface are measured to map the internal structure. On the Earth the waves are observed only after earthquakes, while on the Sun they are continuously excited, probably by the currents in the convective zone. While a wide range of frequencies are observed, the intensity of the oscillation patterns, or modes, peaks strongly at a mode having a period of five minutes. The surface amplitudes range from a few centimetres per second to several metres per second. The modes where the entire Sun expands and contracts or where sound waves travel deeply through the Sun, only touching the surface in a few nodes (*i.e.*, points of no vibration), make it possible to map the deep Sun. Modes with many nodes are, by contrast, limited to the outer regions. Every mode has a definite frequency determined by the structure of the Sun. From a compilation of thousands of mode frequencies, one can develop an independent solar model, which reproduces the observed oscillations quite well. The frequencies of the modes vary slightly with the sunspot cycle.

As the Sun rotates, one half is moving toward us, and the other away. This produces a splitting in the frequencies of the modes (owing to the Doppler shift from the two halves of the Sun). Because the different modes reach different depths in the Sun, the rotation at different depths can be mapped. The rotation of the Sun as a function of depth and latitude is shown in Figure 2. The interior below the convective zone rotates as a solid body. At the surface rotation is fastest at the equator and slowest at the poles. This differential rotation is easily visible as sunspots rotate across the solar surface, and it has been known since the first telescopic studies. At the equator the sunspots rotate at a 25-day rate, and at high latitudes at a 28- or 29-day rate. The differential rotation, apparently generated by the convective zone, is thought to play an important role in the generation of the magnetic field of the Sun. Much is not understood, however, for many solar features show less differential rotation.

Differential rotation

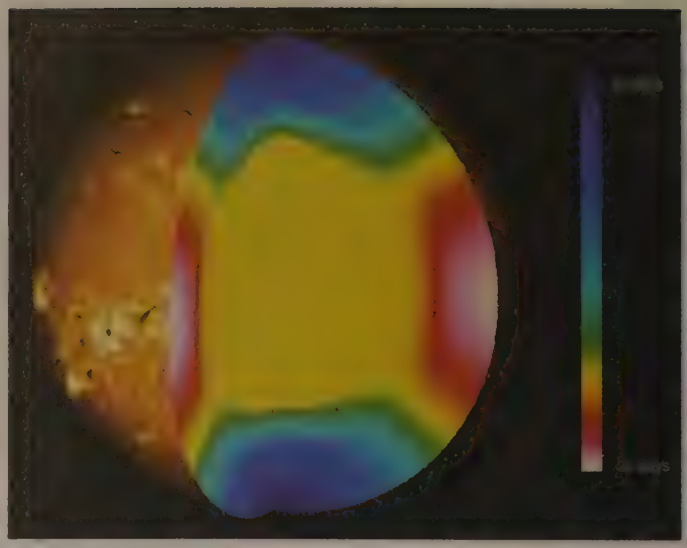


Figure 2: The internal rotation of the Sun as a function of depth and latitude, as derived from helioseismological studies. The differential rotation is clearly shown by the red (fast) area at the equator, extending through the hydrogen convective zone.
Big Bear Solar Observatory, California Institute of Technology

Solar atmosphere

PHOTOSPHERE

Although there are no fires on the surface of the Sun, the photosphere seethes and roils, displaying the effects of the underlying convection. Photons flowing from below, trapped by the underlying layers, finally escape. This produces a dramatic drop in temperature and density. The temperature at the visible surface is about 6,000 K but drops to a minimum about 4,000 K at approximately 500 kilometres above the photosphere. The density, about 10^{-7} gram per cubic centimetre (g/cm^3), drops a factor of 2.7 every 150 kilometres. The solar atmosphere is actually a vacuum by most standards; the total density above any square centimetre is about 1 gram, about 1,000 times less than the comparable mass in the atmosphere of the Earth. One can see through the atmosphere of the Earth but not through that of the Sun because the former is shallow, and the molecules absorb only radiation that lies outside of the visible spectrum. The hot photosphere of the Sun, by contrast, contains an ion called negative hydrogen, H^-, a hydrogen nucleus with two electrons attached. The H^- ion absorbs radiation voraciously through most of the spectrum.

The photosphere is the portion of the Sun seen in ordinary light. Its image reveals two dominant features, a darkening toward the outermost regions, called limb darkening, and a fine rice-grain-like structure called granulation. The darkening occurs simply because the temperature is falling; when one looks at the edge of the Sun, light from higher, cooler, and darker layers is seen. The granules are convective cells that bring energy up from below. Each cell measures about 1,500 kilometres across. Granules have a lifetime of about 25 minutes, during which hot gas rises within them at speeds of about 300 metres per second. They then break up, either by fading out or by exploding into an expanding ring of granules. The granules occur all across the Sun. It is believed that the explosion pattern shapes the surrounding granules in a pattern called mesogranulation, although such a pattern is still in dispute. A larger, undisputed patterned called supergranulation is a network of outward velocity flows, each about 30,000 kilometres across, which is probably tied to the big convective zone rather than to the relatively small granules. The flow concentrates the surface magnetic fields to the supergranulation-cell boundaries, creating a network of magnetic-field elements.

The photospheric magnetic fields extend up into the atmosphere, where the supergranular pattern dominates the conducting gas. While the temperature above the average surface areas continues to drop, it does not fall as rapidly at the network edges, and a picture of the Sun at a wavelength absorbed somewhat above the surface shows the network edges to be bright. This occurs in almost all wavelengths outside the visible. Figure 3 compares a photograph in the near ultraviolet region of the electromagnetic spectrum with a magnetogram of the same area. The ultraviolet emission from the high photosphere closely matches the network of enhanced magnetic field.

Fraunhofer was the first to observe the solar spectrum, finding emission in all colours with many dark lines at certain wavelengths. He assigned letters to these lines, by which some are still known, such as the D-lines of sodium, the G-band, and the K-lines of ionized calcium. Further studies by the German physicist Gustav R. Kirchhoff led

Solar spectrum

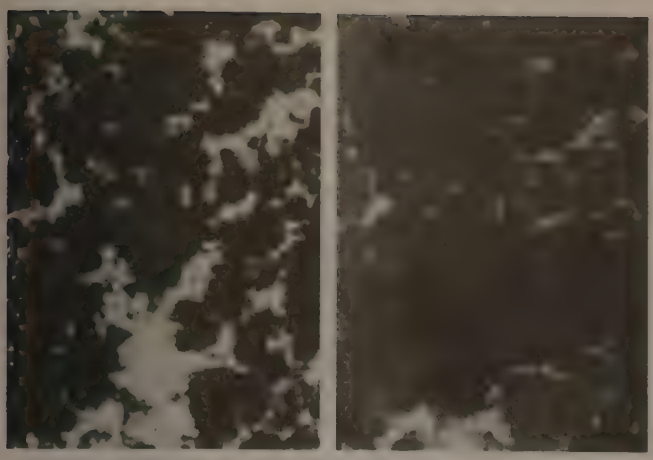


Figure 3: (Left) Image near an active region in the continuum at 1600 angstroms, showing the brightness associated with the magnetic network. (Right) Videomagnetogram of the same region. The fields match the brighter regions almost exactly.

(Left) Laboratoire d'Astrophysique Spatiale, Verneres-le-Buisson, Fr.; (right) Big Bear Solar Observatory, California Institute of Technology

to the understanding that the lines reveal which atoms are in the photosphere and, by comparison with laboratory data, their state of ionization and excitation.

The spectral lines seen are those expected to be common at 6,000 K, where the thermal energy of each particle is about 0.5 volt. The most abundant elements, hydrogen and helium, are difficult to excite, while atoms such as iron, sodium, and calcium have many lines easily excited at this temperature. When Cecilia Payne, a British-born graduate student studying at Harvard College Observatory in Cambridge, Mass., U.S., recognized the great abundance of hydrogen and helium in 1925, she was persuaded by her elders to mark the result as spurious; only later was the truth recognized. The strongest lines in the visible spectrum are the H- and K- (Fraunhofer's letters) lines of ionized calcium. Such is the case because calcium is easily ionized, and these lines represent transitions in which energy is absorbed by ions in the ground, or lowest energy, state. The sodium D-lines are quite a bit weaker because most of the sodium is ionized and does not absorb radiation.

The intensity of the lines is determined by both the abundance of the particular element and its state of ionization, as well as by the excitation of the atomic energy level involved in the line. By working backward one can obtain the abundance of most of the elements in the Sun. This set of abundances occurs with great regularity throughout the universe; it is found in such diverse objects as quasars, meteorites, and new stars. The Sun is roughly 90 percent hydrogen by number of atoms and 9.9 percent helium. The remaining atoms consist of heavier elements, especially carbon, nitrogen, oxygen, magnesium, silicon, and iron, making up only 0.1 percent by number.

CHROMOSPHERE AND CORONA

The ordinary solar spectrum is produced by the photosphere; during an eclipse the brilliant photosphere is blocked out by the Moon and three objects are visible: (1) a thin, pink ring around the edge of the Sun called the chromosphere, (2) a pearly, faint halo extending a great distance, known as the corona, and (3) pink clouds of gas called prominences suspended above the surface. When flash spectra (spectra of the atmosphere during an eclipse) were first obtained, astronomers found several surprising features. First, instead of absorption lines they saw emission lines (bright lines with nothing between them). This effect arises because between the spectrum lines the chromosphere is transparent, and only the dark sky is seen. Second, they discovered that the strongest lines were due to hydrogen, yet they still did not appreciate its high abundance. Finally, the next brightest lines had never been seen before; because they came from the Sun, the unknown source element came to be called helium. Later helium was found on Earth.

Chromosphere. The chromosphere represents the dynamic transition between the cool temperature minimum of the outer photosphere and the diffuse million-degree corona above. It derives its name and pink colour from the red Hα line of hydrogen at 6562.8 angstroms (Å); 1 Å = 10^{-10} metre. Because this line is so strong, it is the best means for studying the chromosphere. For this reason special monochromators are widely used to study the Sun in a narrow wavelength band. Because density decreases with height more rapidly than magnetic field strength, the magnetic field dominates the chromospheric structure, which reflects the extension of the photospheric magnetic fields. The rules for this interplay are simple: every point in the chromosphere where the magnetic field is strong and vertical is hot and hence bright, and every place where it is horizontal is dark. Supergranulation, which concentrates the magnetic field on its edges, produces a chromospheric network of bright regions of enhanced magnetic fields.

The most prominent structures in the chromosphere, especially in the limb, are the clusters of jets, or streams, of particles called spicules (see Figure 4). Spicules extend up to 7,000 kilometres above the surface of the Sun. Because it strongly emits the high-excitation lines of helium, the chromosphere was originally thought to be hot. But radio measurements, a particularly accurate means of measuring the temperature, show it to be only 8,000 K, somewhat hotter than the photosphere. Detailed radio maps show that hotter regions coincide with stronger magnetic fields. Both hot and cold regions extend much higher than one might expect, tossed high above the surface by magnetic and convective action.

Spicules

Big Bear Solar Observatory, California Institute of Technology

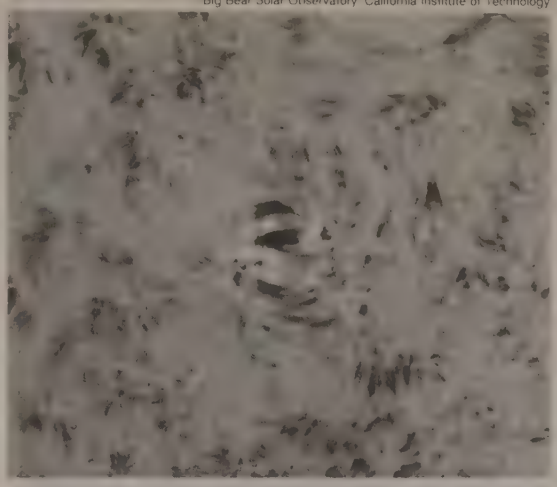


Figure 4: An emerging region of magnetic field in the Hα line, showing the rising arches corresponding to the new flux loops. The dark jets, or streams of particles, around the edges of the network are spicules.

When astronomers observe the Sun from space at ultraviolet wavelengths, the chromosphere is found to emit lines formed at high temperatures, spanning the range from 10,000 to 1,000,000 K. The whole range of ionization of an atom can be found: for example, oxygen I (neutral) is found in the photosphere, oxygen II through VI (one to five electrons removed) in the chromosphere, and oxygen VII and VIII in the corona. This entire series occurs in a height range of about 5,000 kilometres. An image of the corona obtained at ultraviolet wavelengths has a much more diffuse appearance as compared with lower temperature regions, suggesting that the hot material in the magnetic elements spreads outward with height to occupy the entire coronal space. Interestingly, the emission of helium, which was the original clue that the temperature increased upward, is not patchy but uniform. This occurs because the helium atoms are excited by the more diffuse and uniform X-ray emission from the hot corona.

The structure of the chromosphere changes drastically with local magnetic conditions. At the network edges, clusters of spicules project from the clumps of magnetic field lines. Around sunspots, larger field clumps called

plage occur (see below), where there are no spicules, but where the chromosphere is generally hotter and denser. In the areas of prominences the magnetic field lines are horizontal and spicules are absent.

Corona. Another important set of unknown lines revealed during an eclipse came from the corona, and so its source element was called coronium. In 1940 the source of the lines was identified as weak magnetic dipole transitions in various highly ionized atoms such as iron X (iron with nine electrons missing), Fe XIV, and calcium XV, which can exist only if the coronal temperature is about 1,000,000 K. These lines can only be emitted in a high vacuum. The strongest are from iron, which alerted investigators to its high abundance, nearly equal to that of oxygen. Later it was found that there had been errors in prior photospheric determinations.

While the corona is one million times fainter than the photosphere in visible light (about the same as the full Moon at its base and much fainter at greater heights), its high temperature makes it a powerful source of extreme ultraviolet and X-ray emission (see Figure 5). Loops of bright material connect distant magnetic fields. There are regions of little or no corona called coronal holes. The brightest regions are the active regions surrounding sunspots. Hydrogen and helium are entirely ionized, and the other atoms are highly ionized. The ultraviolet portion of the spectrum is filled with strong spectral lines of the highly charged ions. The density at the base of the corona is about 4×10^8 atoms per cubic centimetre, 10^{13} times more tenuous than the atmosphere of the Earth at its base. Because the temperature is high, the density drops by a factor of e (2.718) every 50,000 kilometres.

Coronal holes

Courtesy American Science and Engineering, Cambridge, Massachusetts/National Aeronautics and Space Administration

Figure 5: An X-ray photograph of the Sun taken from the Skylab spacecraft. A great coronal hole is shown crossing the surface.

Radio telescopes are particularly valuable for studying the corona because radio waves will propagate only when their frequency exceeds the so-called plasma frequency of the medium in which they travel. The plasma frequency varies according to the density, temperature, and magnetic fields of the medium, and so measurements of the plasma frequency can yield information about those quantities.

Solar wind. The conductivity of a hot ionized plasma is extremely high, and the coronal temperature decreases only as the $^2/_7$ power of the distance from the Sun. Thus, the temperature of the interplanetary medium is still more than 200,000 K near the Earth. While the gravitational force of the Sun can hold the hot material near the surface, at a distance of $5R_\odot$ the gravitational force is 25 times less, but the temperature is only 40 percent less. Therefore, a continuous outflow of particles known as the solar wind occurs, except where hindered by magnetic fields. The solar wind flows along a spiral path dictated by magnetic fields carried out from the Sun into the interplanetary medium. The velocity is typically 400 kilometres per second, with sizable variation.

Where magnetic fields are strong, the coronal material cannot flow outward and becomes trapped; thus the high density and temperature above active regions is due partly to trapping and partly to heating processes, mostly solar flares. Where the magnetic field is open, the hot material escapes, and a coronal hole results. Analysis of solar wind data shows that coronal holes at the equator are associated with high-velocity streams in the solar wind, and recurrent geomagnetic storms are associated with the return of these holes.

The solar wind drags magnetic field lines out from the surface. Traveling at a speed of 500 kilometres per second, particles will reach the orbit of Saturn in one solar rotation—27 days—but in that time period the source on the Sun will have gone completely around. In other words, the magnetic field lines emanating from the Sun describe a spiral. It takes four days for the solar wind to arrive at the Earth, having originated from a point that has rotated about 50° from its original position facing the Earth. The magnetic field lines, which do not break, maintain the path, and the gas moves along it. The solar-wind flow has a continual effect on the upper atmosphere of the Earth. The total mass, magnetic field, and angular momentum carried away by the solar wind is insignificant, even over the lifetime of the Sun. A higher level of activity in the past, however, might have played a role in the Sun's evolution, and stars larger than the Sun are known to lose considerable mass through such processes.

Since the discovery of the nature of the corona, such low-density, super-hot plasmas have been identified throughout the universe: in the atmospheres of other stars, in supernova remnants, and in the outer reaches of galaxies. Apparently low-density plasmas radiate so little that they can reach and maintain high temperatures. By detecting helium absorption or X-ray emission in stars like the Sun, researchers have found that coronas are quite common. Many stars have coronas far more extensive than that of the Sun.

Coronas of other stars

It is speculated that the high coronal temperature results from boundary effects connected with the steeply decreasing density at the solar surface and the convective currents beneath it. Stars without convective activity do not exhibit coronas. The magnetic fields facilitate a "crack-of-the-whip" effect, in which the energy of many particles is concentrated in progressively smaller numbers of ions. The result is the production of the high temperature of the corona.

Solar activity

SUNSPOTS

A wonderful rhythm in the ebb and flow of sunspot activity dominates the atmosphere of the Sun and influences life on Earth as well. Sunspots, the largest of which can be seen even without a telescope, are regions of extremely strong magnetic field found on the Sun's surface. A typical mature sunspot is seen in white light (Figure 6) to have roughly the form of a daisy. It consists of a dark central core known as the umbra, where the magnetic flux loop emerges vertically from below, surrounded by a ring of dark fibrils called the penumbra, where the magnetic field spreads outward.

The umbra appears dark because it is quite cool, only about 3,000 K, as compared with the 6,000 K temperature of the surrounding photosphere. The spot pressure, consisting of magnetic and gas pressure, must balance the pressure of its surroundings; hence the spot must somehow cool by magnetic effects until the inside gas pressure is considerably lower than that of the outside. Owing to the great magnetic energy present in sunspots, regions near the cool spots actually have the hottest and most intense activity. Sunspots are thought to be cooled by the interference of their strong fields with the convective motions bringing heat from below. For this reason, there appears to be a lower limit on the size of the spots of approxi-

Figure 6: *A sunspot photographed in infrared light.* The dark centre is the umbra, the radiating fibrils are the penumbra, and the rice-grain-like structure surrounding the spot is the granulation. The inner penumbra steadily flows into the umbra.

Big Bear Solar Observatory, California Institute of Technology

mately 500 kilometres. Smaller ones are rapidly heated by radiation from the surroundings and destroyed.

Although the magnetic field suppresses convection and random motions are much lower than in the surroundings, a wide variety of organized motions occur in spots, mostly in the penumbra, where the horizontal field lines permit detectable horizontal flows. One such motion is the Evershed effect, an outward flow at a rate of one kilometre per second in the outer half of the penumbra that extends beyond the penumbra in the form of moving magnetic features. These features are elements of the magnetic field that flow outward across the area surrounding the spot. In the chromosphere above a sunspot, a reverse Evershed flow appears as material spirals into the spot; the inner half of the penumbra flows inward to the umbra.

Evershed effect

Oscillations are observed in sunspots as well. When a section of the photosphere known as a light bridge crosses the umbra, rapid horizontal flow is seen. Although the umbral field is too strong to permit motion, rapid oscillations called umbral flashes appear in the chromosphere just above, with a 150-second period. In the chromosphere above the penumbra, so-called running waves are observed to travel radially outward with a 300-second period.

Magnetic poles, unlike positive protons or negative electrons, cannot exist singly in nature; an isolated magnetic pole could be formed only at enormous energies, and one has never been detected. Therefore, each magnetic pole seen on the Sun has a counterpart of opposite sign, although the two poles may be located far apart. Most frequently, sunspots are likewise seen in pairs, or in paired groups of opposite polarity, which correspond to clusters of magnetic flux loops intersecting the surface. The flux loops emerge from below in pairs of opposite polarity connected by dark arches in the chromosphere above.

The members of a spot pair are identified by their position in the pair with respect to the rotation of the Sun; one is designated as the leading spot and the other as the following spot. In a given hemisphere (north or south), all spot pairs typically have the same polar configuration—*e.g.*, all leading spots may have northern polarity, while all following spots have southern polarity (see below). A new spot group generally has the proper polarity configuration for the hemisphere in which it forms; if not, it usually dies out quickly. Occasionally, regions of reversed polarity survive to grow into large, highly active spot groups. An ensemble of sunspots, the surrounding bright chromosphere, and the associated strong magnetic field regions constitute what is termed an active region. Areas of strong magnetic fields that do not coalesce into sunspots form regions called plages, which are prominent in the red Hα line and are also visible in continuous light near the limb.

Solar activity tends to occur across the entire surface of the Sun at the same time, supporting the idea that the phenomenon is global. While there are sizable variations in the progress of the activity cycle, overall it is impressively regular, indicating a well-established order in the numbers and latitudinal positions of the spots. At the start of a cycle, the number of groups and their size increases rapidly until a maximum in number (known as sunspot maximum) occurs after about two or three years and a maximum in spot area about one year later. The average lifetime of an individual spot group is roughly one solar rotation. The largest spot groups and the greatest eruptions usually occur two or three years after the maximum of the sunspot number. At maximum there might be 10 groups and 300 spots across the Sun, but a huge spot group can have 200 spots in it. The progress of the cycle may be irregular; even near the maximum the number may temporarily drop to low values.

Sunspot maximum

The sunspot cycle returns to a minimum after approximately 11 years. At sunspot minimum there are at most a few small spots on the Sun, usually at low latitudes. New-cycle spots begin to emerge at higher latitudes, between 25° and 40°. These are small and last only a few days. Since the rotation period is 27 days (longer at higher latitudes), these spots usually do not return, and newer spots appear closer to the equator. For a given 11-year cycle, the magnetic polarity configuration of the spot groups is the same in a given hemisphere and is reversed in the opposite hemisphere. The magnetic polarity configuration in each hemisphere reverses in the next cycle. Thus, new spots at high latitudes in the northern hemisphere may have positive polarity leading and negative following, while the groups from the previous cycle, at low latitude, will have the opposite orientation. As the cycle proceeds, the old spots disappear, and new-cycle spots appear in larger numbers at successively lower latitudes. The latitude distribution of spots during a given cycle occurs in a butterfly-like pattern called the butterfly diagram.

The magnetic polarity configuration of the sunspot groups reverses every 11 years; thus, it returns to the same value every 22 years, and this length is considered to be the period of a complete magnetic cycle. At the beginning of each 11-year cycle, the overall solar field, as determined by the dominant field at the pole, has the same polarity as the following spots of the previous cycle. As active regions are broken apart, the magnetic flux is separated into regions of positive and negative sign. After many spots have emerged and died out in the same general area, large unipolar regions of one polarity or the other appear and move toward the Sun's corresponding pole. Owing to the differential rotation of the Sun, the fields approaching the poles rotate more slowly than the sunspots, which at this point in the cycle have congregated in the rapidly rotating equatorial region. Eventually the weak fields reach the pole and reverse the dominant field there. This reverses the polarity to be taken by the leading spots of the new spot groups, thereby continuing the 22-year cycle.

While the sunspot cycle has been quite regular for some centuries, there have been sizable variations. In the period 1955–70 there were far more spots in the northern hemisphere, while in the 1990 cycle they dominated in the southern hemisphere. The two cycles that peaked in 1946 and 1957 were the largest in history. The English astronomer E.W. Maunder found evidence for a period of low activity, pointing out that very few spots were seen between 1645 and 1715. Although sunspots had been first detected about 1600, there are few records of spot sightings during this period, which is called the Maunder minimum. Experienced observers reported the occurrence of a new spot group as a great event, mentioning that they had seen none for years. After 1715 the spots returned. This period was associated with the long cold spell in Europe that extended from about 1550 to 1850 and is known as the Little Ice Age, although it is not proven that this cold period was actually caused by the Maunder

Maunder minimum

minimum. There is evidence for other such low-activity periods at roughly 500-year intervals. When solar activity is high, the strong magnetic fields carried outward by the solar wind block out the high-energy galactic cosmic rays approaching the Earth and less carbon-14 is produced. Measurement of carbon-14 in dated tree rings confirms the low activity at this time. Still, the 11-year cycle was not detected until the 1840s, so observations prior to that time were somewhat irregular.

The origin of the sunspot cycle is not known. Because there is no reason that a star in radiative equilibrium should produce such fields, it is reasoned that relative motions in the Sun twist and enhance magnetic flux loops. The motions in the convective zone may contribute their energy to magnetic fields, but they are too chaotic to produce the regular effects observed. The differential rotation, however, is regular, and it could wind existing field lines in a regular way; hence, most models of the solar dynamo are based on the differential rotation in some respect. The reason for the differential rotation also remains unknown.

Besides sunspots, there exist a great number of tiny spotless dipoles called ephemeral active regions, which last less than a day on average and are found all over the Sun rather than just in the spot latitudes. The number of active regions emerging on the entire Sun is about two per day, while ephemeral regions occur at a rate of about 600 per day. Therefore, even though they are quite small, most of the magnetic flux erupting on the Sun at any one time may be in the form of ephemeral regions. Their small size, however, prohibits them from separating fields and affecting the global field pattern.

PROMINENCES

Prominences are among the most beautiful of solar phenomena. They are the analogues of clouds in the Earth's atmosphere, but they are supported by magnetic fields, rather than by thermal currents as clouds are. Because the plasma of ions and electrons that makes up the solar atmosphere cannot cross magnetic field lines in regions of horizontal magnetic fields, material is supported against gravity. This occurs at the boundaries between one magnetic polarity and its opposite, where the connecting field lines reverse direction. Thus, prominences are reliable indicators of sharp field transitions. Like the chromosphere, prominences are transparent in continuous (white) light and, except during total eclipses, must be viewed in strong Hα spectral lines. The density of prominences is lower than that of the photosphere; they emit less radiation and therefore appear dark. The temperature of prominences is somewhat less than 5,000 K.

Types of prominences

There are two basic types of prominences: (1) quiescent, or long-lived, and (2) transient. The former (Figure 7) are associated with large-scale magnetic fields, marking the boundaries of unipolar magnetic regions or sunspot groups. Because the large unipolar plates are long-lived, the quiescent prominences are as well. They may have varied forms—hedgerows, suspended clouds, or funnels—but they always take the form of two-dimensional suspended sheets. Stable filaments often become unstable and erupt, but they may also just fade away. The prominences never fall downward; they always erupt upward, because all unattached magnetic fields have a tremendous buoyancy and attempt to leave the Sun. When they do escape, they produce not only a splendid sight but also a transient shock wave in the corona called a coronal mass ejection, which can cause important geomagnetic effects.

Transient prominences are also part of solar activity. Sprays are the disorganized mass of material ejected by a flare (see Figure 8). Loop prominences are the aftermath of flares. Surges are collimated streams of ejecta connected with small flares.

The spectrum of quiescent prominences seen against the sky is essentially similar to the chromosphere—*i.e.*, it reflects the properties of a gas excited primarily by photospheric emission. By contrast with the chromosphere, where spicule motions produce broad lines, the lines are quite narrow, indicating little internal motion. Spectra of transient prominences reflect quite different environments. Because they usually are part of a very hot flare or a condensation from the hot corona, they show high-excitation lines of ionized helium and strong ultraviolet emission, as befits a gas at 30,000 to 100,000 K.

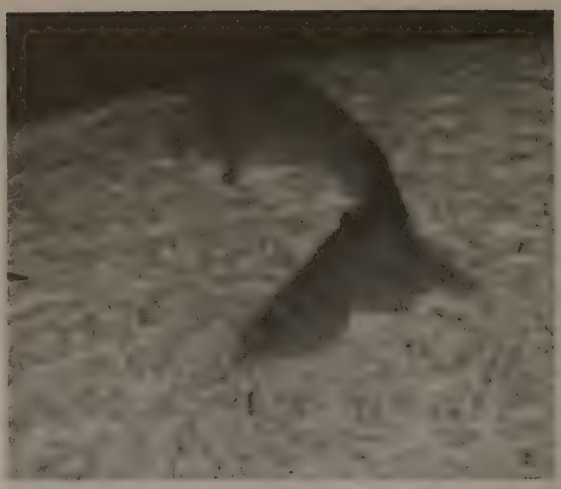


Figure 7: A quiescent prominence, or filament, dark against the Sun's disk, photographed in Hα. Spicules (dark jets) are seen in the chromosphere below.
Big Bear Solar Observatory, California Institute of Technology

FLARES

The most spectacular phenomenon related to sunspot activity is the solar flare, which is an abrupt release of magnetic energy from the sunspot region. Despite the great energy involved, most flares are almost invisible in ordinary light because the energy release takes place in the transparent atmosphere, and only the photosphere, which relatively little energy reaches, can be seen in visible light. Flares are best seen in the Hα line, where the brightness may be 10 times that of the surrounding chromosphere, or 3 times that of the surrounding continuum. In Hα a big flare will cover a few thousandths of the Sun's disk, but in white light only a few small bright spots appear. The energy released in a great flare can reach 10^{33} ergs, which is equal to the output of the entire Sun in 0.25 second. Most of this energy is initially released in high-energy electrons and protons, and the optical emission is a secondary effect caused by the particles impacting the chromosphere.

Classification of flares

There is a wide range of flare size, from giant events that shower the Earth with particles to brightenings that are barely detectable. Flares are usually classified by their associated flux of X rays having wavelengths between one and eight angstroms: C*n*, M*n*, or X*n* for flux greater than 10^{-6}, 10^{-5}, and 10^{-4} watts per square metre (w/m^2), respectively, where the integer n gives the flux for each power of 10. Thus, M3 corresponds to a flux of 3×10^{-5} w/m^2 at the Earth. This index is not linear in flare energy since it measures only the peak, not the total. The energy released in the three or four biggest flares each year is equivalent to the sum of the energies produced in all the small flares. A flare can be likened to a giant natural synchrotron ac-

Big Bear Solar Observatory, California Institute of Technology

Figure 8: *Eruption from a flare just inside the limb.* The upper portion of the photograph has been enhanced to bring out the erupting prominence.

celerating vast numbers of electrons to energies above 10 thousand-electron volts (KeV) and protons to those above one million-electron volts (MeV). Almost all the energy initially goes into these high-energy particles, which subsequently heat the atmosphere or travel into interplanetary space. The electrons produce X-ray bursts and radio bursts and also heat the surface. The protons produce gamma-ray lines by exciting or breaking up surface nuclei. Both electrons and protons propagate to the Earth; the clouds of protons bombard the Earth in big flares. Most of the energy heats the surface and produces a hot (40,000,000 K) and dense cloud of coronal gas, which is the source of the X rays. As this cloud cools, the elegant loop prominences appear and rain down to the surface.

Flares occur along neutral lines—*i.e.*, boundaries between regions of opposite magnetic polarity. While opposite polarities are often stably connected across such boundaries, the field is sheared and distorted when relative motion or rapid flux emergence occurs, and the energy is cataclysmically released in flares. Impulsive flares are accompanied by outward explosion and ejection of material; the material may be carried away with the erupting magnetic field or may be ejected by the high pressure in the flare. The highest recorded speed is 1,500 kilometres per second, but 100–300 kilometres per second is more typical. Great clouds of coronal material are blown out; these make up a substantial fraction of the solar wind.

Most of the great flares occur in a small number of superactive sunspot groups; an example is shown in Figure 9. The groups are characterized by a large cluster of spots of one magnetic polarity surrounded by the opposite polarity. Although the occurrence of flares can be predicted from the presence of such spots, researchers cannot predict when these mighty regions will emerge from below the surface, nor do they know what produces them. A flare exploding from a complex spot is shown in Figure 10.

SOLAR-TERRESTRIAL EFFECTS

Besides providing light and heat, the Sun affects the Earth through its ultraviolet radiation, the steady stream of the solar wind, and the particle storms of great flares. The near-ultraviolet radiation from the Sun produces the ozone layer, which in turn shields the planet from such radiation. The soft (long-wavelength) X rays from the solar corona produce those layers of the ionosphere that make short-wave radio communication possible. The harder (shorter-wavelength) X-ray pulses from flares ionize the lowest ionospheric layer, producing radio fadeouts. The Earth's rotating magnetic field is strong enough to block the solar wind, forming the magnetosphere, around which the solar particles and fields flow. On the side opposite to the Sun, the field lines stretch out in a structure called the magnetotail. When shocks arrive in the solar wind, a short, sharp increase in the field of the Earth is produced. When the interplanetary field switches to a direction opposite the Earth's field, or when big clouds of particles enter it, the fields in the magnetotail reconnect and energy is released, producing the aurora borealis (northern lights). Big flares or coronal mass ejections bring clouds of energetic particles that form a ring current around the magnetosphere, which produces sharp fluctuations in the Earth's field called geomagnetic storms. These phenomena disturb radio communication and produce voltage surges in long-distance transmission lines and other long conductors.

Magnetic field effects

Big Bear Solar Observatory, California Institute of Technology

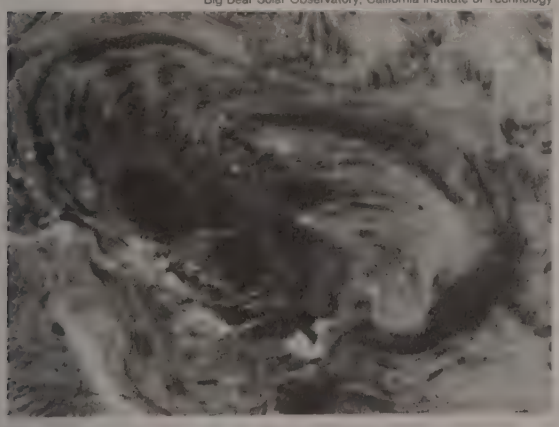


Figure 9: *The superactive spot group of June 1991 in Hα.* The typical "island δ" spot combined large spots of opposite polarity that were closely forced together. The dark loops are overlying prominences.

Figure 10: *A great flare exploding from a complex spot.* A shock front flew across the Sun to the right, and a great X-ray burst occurred.

Big Bear Solar Observatory, California Institute of Technology

Perhaps the most intriguing of all terrestrial effects are the possible effects of the Sun on the climate of the Earth. The Maunder minimum seems well established, but there are few other clear effects. The only other definite relationship is the temperature change associated with the quasi-biennial oscillation of the tropical stratospheric wind. The brightness of the Sun varies with activity; a large sunspot reduces emission by an amount corresponding to its area. The effects of plages produce an overall increase in the solar flux by about 0.1 percent, however, when spot activity increases. This is a negligible effect, and so particle effects and fluctuations of ultraviolet radiation in the stratosphere are thought to be important.

Because charged particles follow magnetic fields, corpuscular radiation is not observed from all big flares but only from those favourably situated in the Sun's western hemisphere. The solar rotation makes the lines of force from the western side of the Sun (as seen from the Earth) lead back to the Earth, guiding the flare particles there. These particles are mostly protons because hydrogen is the dominant constituent of the Sun. Many of the particles are trapped in a great shock front that blows out from the Sun at 1,000 kilometres per second. The flux of low-energy particles in big flares is so intense that it endangers the lives of astronauts outside the terrestrial magnetic field.

History of observation

The existence of features on the Sun was known from the records of sunspots observed by ancient astronomers with the naked eye; however, no systematic studies were made of such features until the telescope was invented in the early 17th century. The Italian scientist Galileo and the German mathematician Christoph Scheiner were among the first to make telescopic observations of sunspots. Scheiner's drawings in the *Rosa Ursina* are of almost modern quality, and there was little improvement in solar imaging until 1905. In the 1670s the British astronomer John Flamsteed and the French astronomer Gian Domenico Cassini calculated the distance to the Sun; using data from observations of the transits of Venus in 1761 and 1769, scientists were able to determine the distance between the Sun and the Earth

more precisely—their estimations were quite close to modern values. Newton set forth the role of the Sun as the centre of attraction of the known planetary system.

While the quality of observations was good, consistent observation was lacking. The sunspot cycle, a huge effect, was not discovered until 1843 by Heinrich Schwabe. The German amateur astronomer was looking for a planet inside the orbit of Mercury and made careful daily drawings to track its passage across the face of the Sun. Instead he found that the number of sunspots varied with a regular period. The Swiss astronomer Rudolf Wolf confirmed Schwabe's discovery by searching through previous reports of sunspots and established the period as 11 years. Wolf also introduced what is known as the Zurich relative sunspot number (or Wolf's sunspot number), a value equal to the sum of the spots plus 10 times the number of groups, which is still used today. Much of the work at this time was carried out by wealthy amateurs such as Richard Christopher Carrington of Britain, who built a private observatory and discovered the differential rotation and the equatorward drift of activity during a sunspot cycle. He also was the first (with another Englishman, R. Hodgson) to observe a solar flare. Photographic monitoring began in 1860, and soon spectroscopy was applied to the Sun, so the elements present and their physical state could begin to be investigated. In the early part of the 19th century, Fraunhofer mapped the solar spectrum. At the end of the 19th century, spectroscopy carried out during eclipses revealed the character of the atmosphere, but the million-degree coronal temperature was not established until 1940 by the German astrophysicist Walter Grotrian.

Discovery of the 11-year cycle

In 1891, while he was a senior at the Massachusetts Institute of Technology in Cambridge, the American astronomer George Ellery Hale invented the spectroheliograph, which can be used to take pictures of the Sun in any wavelength. After using the instrument on the great Yerkes refractor in Williams Bay, Wis. (which he built), Hale developed the Mount Wilson Observatory in California and built the first solar tower telescopes there. Prior to the construction of the Mount Wilson facility, all solar observatories were located in cloudy places, and long-term studies were not possible. Hale discovered the magnetic fields of sunspots by observing the splitting of their spectral lines into a number of components; this splitting, known as the Zeeman effect, occurs in the presence of a strong magnetic field. By continuously studying the spots for two cycles, he discovered, with the American astronomer Seth Barnes Nicholson, the law of sunspot polarities. Later his successors developed the magnetograph, with which the polar field was detected. In the 1930s the French astronomer Bernard Lyot introduced the coronagraph, which made possible spectral observations of the corona when the Sun is not in eclipse, and the birefringent filter, which permitted two-dimensional monochromatic images. With the Lyot filter, cinematography of the solar activity of magnetic and velocity fields became a reality.

Invention of the spectroheliograph

After 1950 new observatories were established in areas that were less cloudy. By 1960 it was realized that these sites not only had to be clear, but they also had to have stable air. By situating observatories near lakes and employing electronic imaging and vacuum telescopes, astronomers were able to make new, higher-resolution observations. It is now possible to scan the surface of the Sun as if flying over it in an airplane, taking digital images of high quality that reveal the solar magnetic and velocity fields.

(H.Zi.)

THE MAJOR PLANETS AND THEIR SATELLITES

Mercury

Mercury, designated ☿ in astronomy, is the innermost planet of the solar system and eighth in size and mass. Its closeness to the Sun and smallness make it the most elusive of the planets visible to the unaided eye. Because its rising or setting is always within about two hours of the Sun's, it is never observable when the sky is fully dark.

The difficulty in seeing it notwithstanding, Mercury was known at least by Sumerian times, some 5,000 years ago. In Classical Greece it was called Apollo when it appeared as a morning star just before sunrise and Hermes, the Greek equivalent of the Roman god Mercury, when it appeared as an evening star just after sunset. Hermes was the swift messenger of the gods, and the planet's name is thus likely a reference to its rapid motions relative to other objects in the sky.

Shortage of basic information

Until the last part of the 20th century, Mercury was one of the least-understood planets, and even now the shortage of information about it leaves many basic questions unsettled. Indeed, the length of its day was not determined until the 1960s, and at the turn of the 21st century the appearance of half of its surface was still unknown. At first glance the hemisphere of the planet that has been imaged looks similar to the cratered terrain of the Moon, an impression reinforced by the roughly comparable size of the two bodies. Mercury is far denser, however, having a metallic core that takes up about 42 percent of its volume (compared with 4 percent for the Moon and 16 percent for Earth). Moreover, its surface shows significant differences from lunar terrain, including a lack of the massive dark-coloured lava flows known as maria on the Moon and the presence of buckles and scarps that suggest Mercury shrank during some period in its history.

BASIC ASTRONOMICAL DATA

Mercury is an extreme planet in several respects. Because of its nearness to the Sun—its average orbital distance is 58 million kilometres—it has the shortest year (a revolution period of 88 days) and receives the most intense solar radiation of all the planets. With a radius of about 2,440 kilometres, Mercury is, except for distant Pluto, the smallest planet, smaller even than Jupiter's largest moon, Ganymede, or Saturn's largest moon, Titan. In addition, Mercury is unusually dense. Although its mean density is roughly that of Earth's, it has less mass and so is less compressed by its own gravity; when corrected for self-compression, Mercury's density is the highest of any planet. Nearly two-thirds of Mercury's mass is contained in its largely iron core, which extends from the planet's centre to a radius of about 1,800 kilometres, or three-quarters of the way to its surface. The planet's rocky outer shell—its surface crust and underlying mantle—is only some 600 kilometres thick. For additional orbital and physical data, see Table 2.

Intrinsically densest planet

Table 2: Planetary Data for Mercury

Mean distance from Sun	57,910,000 km (0.4 AU)
Eccentricity of orbit	0.206
Inclination of orbit to ecliptic	7.004°
Mercurian year (sidereal period of revolution)	87.9694 Earth days
Rotation period (Mercurian sidereal day)	58.6462 Earth days
Mean synodic period	116 Earth days
Mean orbital velocity	48 km/s
Inclination of equator to orbit	less than 3° (probably nearly 0°)
Mass	3.30×10^{23} kg
Radius	2439.7 km
Density	5.43 g/cm³
Mean surface gravity	370 cm/s²
Magnetic field strength	0.003 gauss
Mean surface temperature	440 K (332° F, 167° C)
Number of known moons	none

Observational challenges. As seen from Earth's surface, Mercury hides in dusk and twilight, never getting more than about 28° in angular distance from the Sun. It takes about 116 days for successive elongations—*i.e.*, for Mercury to return to the same point relative to the Sun—in the morning or evening sky; this is called Mercury's synodic period. Its nearness to the horizon also means that Mercury is always seen through more of Earth's turbulent atmosphere, which blurs the view. Even orbiting observa-

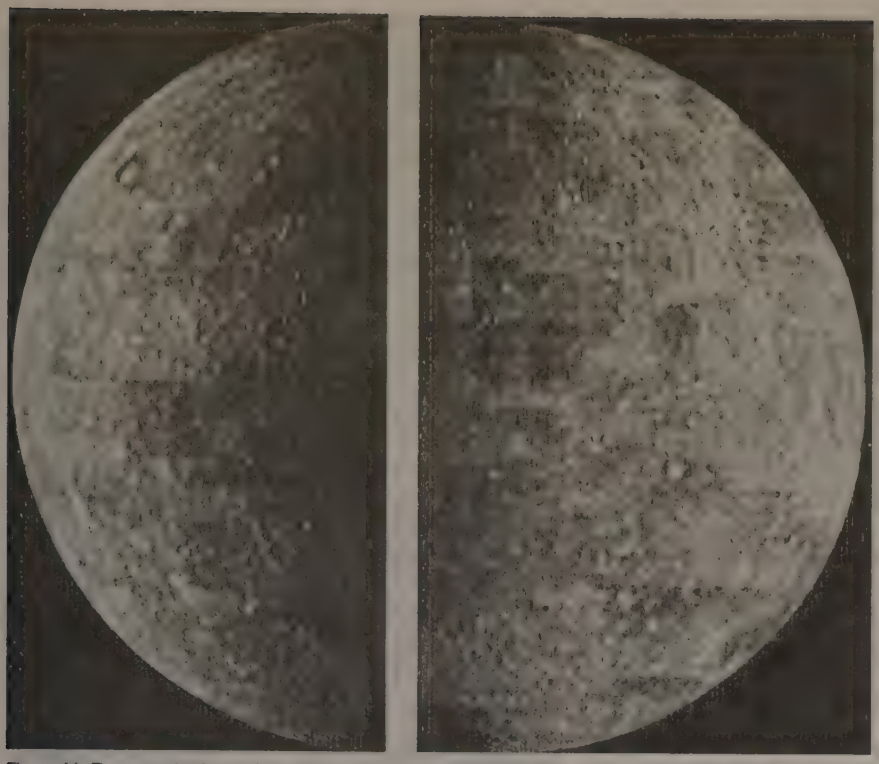


Figure 11: Two mosaic views of Mercury, each showing about half of the hemisphere that was in sunlight when Mariner 10 made its first flyby in March 1974. In each image the landscape is dominated by large impact basins and craters with extensive intercrater plains. Half of the enormous Caloris impact basin appears along the terminator in the right image (left and just above centre).
NASA/JPL

tories such as the Hubble Space Telescope are restricted by the high sensitivity of their instruments from pointing as close to the Sun as would be required for observing Mercury. Because Mercury's orbit lies within Earth's, it occasionally passes directly between Earth and the Sun. This event, in which the planet can be observed telescopically or by spacecraft instruments as a small black dot crossing the bright solar disk, is called a transit, and it occurs about a dozen times in a century.

Transits of the Sun

Mercury also presents difficulties to study by space probe. Because the planet is located deep in the Sun's gravity field, a great deal of energy is needed to shape the trajectory of a spacecraft to get from Earth's orbit to Mercury's such that it can go into orbit around the planet or land on it. The only spacecraft to visit Mercury, Mariner 10, was in orbit around the Sun when it made three brief flybys of the planet in 1974–75. In developing future missions to Mercury, such as the U.S. Messenger spacecraft launched in 2004, spaceflight engineers have calculated complex routes, making use of gravity assists from repeated flybys of Venus and Mercury over several years. In the Messenger mission design, after conducting flyby observations, the spacecraft would enter into an elongated orbit around Mercury for close-up investigations.

Orbital and rotational effects. Mercury's orbit is inclined about 7° to the ecliptic, the plane defined by the orbit of Earth around the Sun; it is also unusually eccentric, or elongated. As a result of the elongated orbit, the Sun appears more than twice as bright in Mercury's sky when the planet is closest to the Sun (at perihelion), at 46 million kilometres, than when it is farthest from the Sun (at aphelion), at nearly 70 million kilometres. The planet's rotation period of 58.6 Earth days with respect to the stars—*i.e.*, the length of its sidereal day—causes the Sun to drift slowly westward in Mercury's sky. Because Mercury is also orbiting the Sun, its rotation and revolution periods combine such that the Sun takes three Mercurian sidereal days, or 176 Earth days, to make a full circuit—the length of its solar day.

As described by Kepler's laws of planetary motion, Mercury travels around the Sun so swiftly near perihelion that the Sun appears to reverse course in Mercury's sky, briefly

NASA/TRACE/SMEX

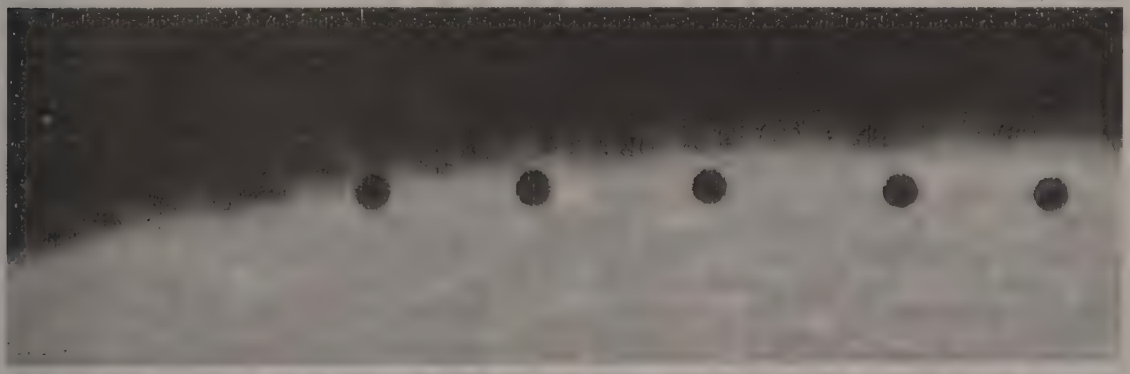


Figure 12: Transit of Mercury across the face of the Sun, Nov. 15, 1999, a composite of images in ultraviolet light taken by the Transition Region and Coronal Explorer (TRACE) satellite in Earth orbit. The time between successive images is about seven minutes.

moving eastward before resuming its westerly advance. The two locations on Mercury's equator where this oscillation takes place at noon are called hot poles. As the overhead Sun lingers there, heating them preferentially, surface temperatures can exceed 700 kelvins (K; 800° F, 430° C). The two equatorial locations 90° from the hot poles, called warm poles, never get nearly as hot. Near the north and south rotational poles of Mercury, ground temperatures are even colder, below 200 K (−100° F, −70° C), when lit by grazing sunlight. Surface temperatures drop to about 90 K (−300° F, −180° C) during Mercury's long nights before sunrise.

Hot and warm poles

Mercury's nightside would be even colder if the planet kept one face perpetually toward the Sun and the other in perpetual darkness. Until Earth-based radar observations in the 1960s proved otherwise, astronomers had long believed that to be the case, which would follow if Mercury's rotation were synchronous—that is, if its rotation period were the same as its 88-day revolution period. The radar studies revealed that the planet's 58.6-day rotation period is not only different from its orbital period but also exactly two-thirds of it.

Mercury's orbital eccentricity and the strong solar tides—deformations raised in the body of the planet by the Sun's gravitational attraction—apparently explain why the planet rotates three times for every two times that it orbits the Sun. Mercury presumably had spun faster when it was forming, but it was slowed by tidal forces. Instead of slowing to a state of synchronous rotation, as has happened to many planetary satellites, including Earth's Moon, Mercury became trapped at the 58.6-day rotation rate. At this rate the Sun tugs repeatedly and especially strongly on the tidally induced bulges in Mercury's crust at the hot poles.

Mariner 10 and later studies. Most of what scientists know about Mercury was learned during the three flybys by Mariner 10. Because the spacecraft was placed in an orbit around the Sun equal to one Mercurian solar day, it made each of its three passes when exactly the same half of the planet was in sunlight. Slightly less than the illuminated half, or about 45 percent of Mercury's surface, was eventually imaged. Mariner 10 also collected data on particles and magnetic fields during its flybys. Mercury was discovered to have a surprisingly Earth-like (though much weaker) magnetic field. Scientists had not anticipated a planetary magnetic field for such a small, slowly rotating body, because the dynamo theories that described the phenomenon required thoroughly molten cores and rather rapid planetary spins. Even more rapidly spinning bodies such the Moon and Mars lack magnetic fields. In addition, Mariner 10's spectral measurements showed that Mercury has an extremely tenuous atmosphere.

Unexpected planetary magnetic field

The first significant telescopic data about Mercury after the Mariner mission resulted in the discovery in the mid-1980s of sodium in the atmosphere. Subsequently, better Earth-based techniques enabled the variations of several of Mercury's atmospheric components to be studied from place to place and over time. Also, improvement in the power and sensitivity of ground-based radar resulted in intriguing maps of the regions unseen by Mariner 10.

THE ATMOSPHERE

A planet as small and as hot as Mercury has no possibility of retaining a significant atmosphere, if it ever had one. To be sure, Mercury's surface pressure is less than one-trillionth that of Earth. Nevertheless, the traces of atmospheric components that have been detected have provided clues about interesting planetary processes. Mariner 10 found small amounts of atomic helium and even smaller amounts of atomic hydrogen near Mercury's surface. These atoms are mostly derived from the solar wind—the flow of charged particles from the Sun that expands outward through the solar system—and remain near Mercury's surface for very short times, perhaps only hours, before escaping the planet. Mariner also detected atomic oxygen, which, along with tiny amounts of sodium and potassium found subsequently in telescopic observations, is probably derived from Mercury's surface soils or impacting meteoroids and ejected into the atmosphere either by the impacts or by bombardment of solar wind particles. The atmospheric gases tend to accumulate on Mercury's nightside but are dissipated by the brilliant morning sunlight. Presumably many other gases, less easy to detect, are present in similar minuscule quantities.

In the early 1990s Earth-based radar made the remarkable discovery of patches of highly radar-reflective materials at the poles, apparently only in permanently shadowed regions of deep, near-polar craters. Scientists believe that the reflecting material might be water ice. The idea that the planet nearest the Sun might harbour significant deposits of water ice originally seemed bizarre. Yet, Mercury must have accumulated water over its history—for example, from impacting comets. Water ice on Mercury's broiling surface will immediately turn to vapour, and the individual water molecules will hop, in some random direction, along ballistic trajectories. Calculations suggest that after many hops perhaps 1 out of 10 water molecules eventually lands in a deep polar depression.

Possibility of polar ice deposits

Courtesy of John Harmon, Arecibo Observatory

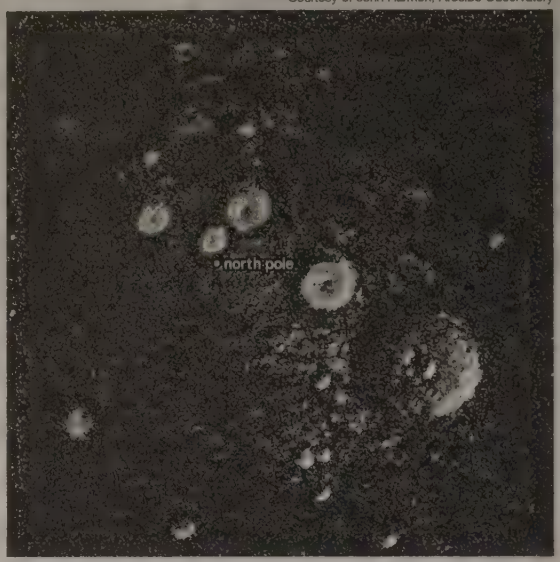


Figure 13: Mercury's north polar region, in a radar image obtained with the Arecibo radio telescope. All the bright (radar-reflective) features are believed to be deposits of frozen volatile substances, likely water ice, in the permanently shaded floors of craters.

Because Mercury's rotational axis is essentially perpendicular to the plane of its orbit, sunlight is always nearly horizontal at the poles. Under such conditions the bottoms of deep depressions would remain in permanent shadow and provide cold traps that could hold water molecules for millions or billions of years. Gradually a polar ice deposit would build up. The susceptibility of the ice to subliming away slowly—*e.g.*, from the slight warmth of sunlight reflected from distant mountains or crater rims—could be reduced if it gradually became cloaked by an insulating debris layer, or regolith, made of dust and rock fragments ejected from distant impacts. Radar data suggest that the reflecting layer indeed is covered with as much as 0.5 metre of such debris.

It is far from certain that the volatile material near Mercury's poles is water ice. Additional radar studies found small patches of high reflectivity at latitudes as low as 71°, where water ice would be far less likely to form and survive. Moreover, the same reasoning about the possibility of water ice near Mercury's poles also has been applied to the Moon, where the accumulation process should have been even more robust. In its orbital mission in 1998–99, the Lunar Prospector spacecraft found evidence for, at most, minimal water ice near the lunar poles. Perhaps another easily evaporated substance, but one less volatile than water, has been "cold-trapped" on Mercury.

THE MAGNETIC FIELD AND MAGNETOSPHERE

As closely as Mariner 10's measurements could determine, Mercury's magnetic field, though only 1 percent as strong as Earth's, resembles Earth's field in being roughly dipolar

Figure 14: Mercury's Caloris impact basin, in a mosaic of images captured by Mariner 10 during its three flybys. Only the eastern half is visible; it appears as partial concentric rings (left portion of image) within relatively smooth plains.
NASA/JPL

and oriented along the planet's spin axis. While the existence of the field might conceivably have another explanation—for example, remanent magnetism, the retained imprint of an ancient magnetic field frozen into the rocks during crustal cooling—most researchers are convinced that it is produced, like Earth's field, by a magnetohydrodynamic dynamo mechanism involving motions within an electrically conducting fluid core.

Mercury's magnetic field holds off the solar wind with a teardrop-shaped bubble, or magnetosphere, whose rounded end extends outward toward the Sun about one planetary radius from the surface. This is only about 5 percent of the sunward extent of Earth's magnetosphere. The planet's atmosphere is so thin that no equivalent to Earth's ionosphere exists at Mercury. Indeed, calculations suggest that on rare occasions the solar wind is strong enough to push the sunward boundary (magnetopause) of the magnetosphere beneath Mercury's surface. Under these conditions solar wind ions would impinge directly on the portions of Mercury's surface immediately beneath the Sun. Even infrequent occurrences of this event could dramatically alter the atomic composition of surface constituents. Mercury's magnetospheric processes are of interest to geophysicists and space scientists, who hope one day to test their conception of Earth's magnetosphere through examination of an Earth-like field with a very different scale and in a different solar wind environment.

Impingement of solar wind on surface

CHARACTER OF THE SURFACE

The portion of Mercury imaged by Mariner 10 looks, superficially, like the Moon. Mercury is heavily pockmarked with impact craters of all sizes. The smallest craters visible in the highest-resolution Mariner photos are a few hundred metres in diameter. Interspersed among the craters are relatively flat, less-cratered regions termed intercrater plains. These are similar to but much more pervasive than the light-coloured plains that occupy intercrater areas on the heavily cratered highlands of the Moon. There are also some sparsely cratered regions called smooth plains, many of which surround the most prominent impact structure on Mercury, the immense impact basin known as Caloris, only half of which was in sunlight during the Mariner encounters.

Impact craters. The most common topographic features on Mercury are the craters that cover much of its surface. Although lunarlike in general appearance, Mercurian craters show interesting differences when studied in detail. Mercury's surface gravity is more than twice that of the Moon, partly because of the great density of the planet's huge iron core. The higher gravity tends to keep material ejected from a crater from traveling as far—only 65 percent of the distance that would be reached on the Moon. It also means that the complex forms and structures characteristic of larger craters—central peaks, slumped crater walls, and flattened floors—occur in smaller craters on Mercury (minimum diameters of about 10 kilometres) than on the Moon (about 19 kilometres). Craters smaller than these minimums have simple bowl shapes.

Comparisons with craters on Moon and Mars

Mercury's craters also show differences from those on Mars, although the two planets have comparable surface gravities. Fresh craters tend to be deeper on Mercury than craters of the same size on Mars; this may be because of a lower content of volatile materials in the Mercurian crust or higher impact velocities on Mercury (the velocity of an object in solar orbit increases with nearness to the Sun).

Craters on Mercury larger than about 100 kilometres in diameter begin to show features indicative of a transition to the "bull's-eye" form that is the hallmark of the largest impact basins. These latter structures, called multiring basins and measuring 300 kilometres or more across, are products of the most energetic impacts. About two dozen multiring basins have been recognized on the imaged portion of Mercury; two or three exceed the size of Caloris, although they are older and less prominent.

Caloris and its antipodal region. The ramparts of the Caloris impact basin span a diameter exceeding 1,300 kilometres. Its interior is occupied by smooth plains that are extensively ridged and fractured in a crudely radial and concentric pattern. The largest ridges are a few hundred kilometres long, about 3 kilometres wide, and less than 300 metres high. Fractures are comparable to ridges in size, and some resemble depressions bounded by faults (grabens). Where they cross ridges, they cut through them.

Two types of terrain surround Caloris—the basin rim and the basin ejecta terrains. The rim consists of a ring of irregular mountain blocks approaching 3 kilometres in height, the highest mountains yet seen on Mercury, bounded on the interior by a relatively steep slope, or escarpment. A second, much smaller escarpment ring stands about 100–150 kilometres beyond the first. Smooth plains occupy the depressions between mountain blocks. Beyond the outer escarpment is a zone of linear, radial ridges and valleys that are partially filled by plains, some with numerous knobs and hills only a few hundred metres across.

Caloris is the youngest of the large multiring basins, at least on the observed side of Mercury. It probably was formed at the same time as the last giant basins on the Moon, but possibly more recently.

On the other side of the planet, exactly 180° opposite Caloris, is a region of weirdly contorted terrain (Figure 15). It is thought to have been formed at the same time as the Caloris impact by the focusing of seismic waves from that event to the antipodal area on Mercury's surface. Termed hilly and lineated terrain, it is an extensive area of elevations and depressions. The crudely polygonal hills are 5–10 kilometres wide and up to 1.5 kilometres high.

Region opposite Caloris

Plains. Plains—relatively flat or smoothly undulating surfaces—are ubiquitous on Mercury and the other terrestrial planets. They represent a canvas on which other landforms develop. The covering or destruction of a rough topography and the creation of a smoother surface is called resurfacing, and plains are evidence of this process.

There are at least three ways that planets are resurfaced, and all three may have had a role in creating Mercury's plains. One way, raising the temperature, reduces the strength of the crust and its ability to retain high relief; over millions of years the mountains sink and the crater basins rise. A second way involves the flow of material toward lower elevations under the influence of gravity; the material eventually collects in depressions and fills to higher levels as more volume is added. Flows of lava from the interior behave in this manner. A third way is for fragments of material to be deposited on a surface from above, first mantling and eventually obliterating the rough topography. Blanketing by impact crater ejecta and volcanic ash is an example of this mechanism.

NASA/JPL

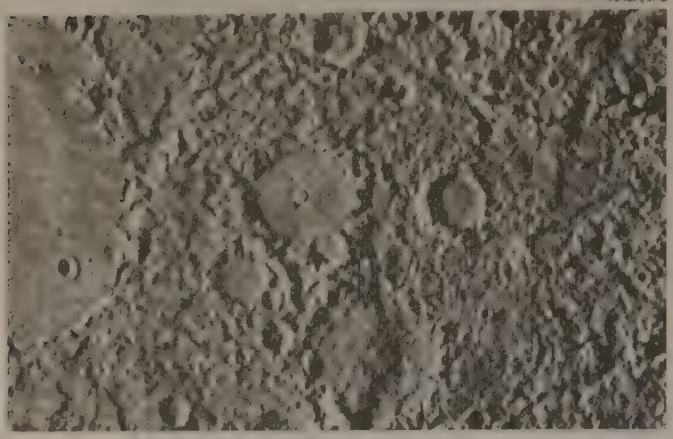


Figure 15: Hilly and lineated terrain located diametrically across Mercury from Caloris, imaged by Mariner 10 in 1974.

Composition of intercrater plains

Even though the Moon has always been much more accessible than Mercury to remote observation, it took the elaborate manned Apollo investigations to decide on the origin of the lunar intercrater plains. The available information about Mercury is still inadequate to conclude decisively whether its widespread intercrater plains are composed of materials ejected from ancient large impacts, as they have been determined to be on the Moon, or instead are volcanic lava flows. Recent interpretations of Mariner 10 images favour volcanic outpourings as the origin of many of the smooth plains on Mercury, especially the smooth plains near and within Caloris. On the other hand, few if any topographic features characteristic of volcanic activity (*e.g.*, solidified lava flow fronts) have been found, although that is partly because of the relatively coarse quality of the Mariner images.

Scarps. The most important landforms on Mercury for gaining insight into the planet's otherwise largely unseen interior workings have been its hundreds of lobate scarps. These cliffs vary from tens to hundreds of kilometres in length and from about 100 metres to 3 kilometres in altitude. Viewed from above, they have curved or scalloped edges, hence the term *lobate*. It is clear that they were formed from fracturing, or faulting, when one portion of the surface was thrust up and overrode the adjacent terrain. On Earth such thrust faults are limited in extent and result from local horizontal compressive (squeezing) forces in the crust. On Mercury, however, these features range across the 45 percent of the surface that has been imaged, which implies that Mercury's crust must have contracted globally. From the numbers and geometries of the lobate scarps, it appears that the planet shrank in diameter by an astonishing 2 kilometres.

Global shrinkage

Moreover, the shrinkage must have occurred comparatively recently in Mercury's geologic history, about the time of the formation of Caloris, because the lobate scarps have been altered only by the most recent activity associated with the Caloris impact and by the freshest-appearing impact craters. The slowing of the planet's initial high rotation rate by tidal forces (see above *Orbital and rotational effects*) would have produced compression in Mercury's equatorial latitudes. The globally distributed lobate scarps, however, suggest another explanation: later cooling of the planet's mantle, perhaps combined with freezing of part of its once totally molten core, caused the interior to shrink and the cold surface crust to buckle. In fact, the contraction of Mercury estimated from cooling of its mantle should have produced even more compressional features on its surface than have been seen, which suggests that the planet has not finished shrinking.

Surface composition. Scientists have attempted to make deductions about the makeup of Mercury's surface from studies of the sunlight reflected from different regions. Among the differences noted between Mercury and the Moon are that the range of surface brightnesses is narrower on Mercury and that colour differences across the planet are less pronounced. These attributes, as well as the relatively featureless visible and near-infrared spectrum of its reflected sunlight, suggest that Mercury's surface is lacking in iron- and titanium-rich silicate minerals compared with the lunar maria. In particular, Mercury's rocks may be low in oxidized iron (FeO), and this leads to speculation that the planet was formed in conditions much more reducing—*i.e.*, those in which oxygen was scarce—than other terrestrial planets.

Determination of the composition of Mercury's surface from remote-sensing data is fraught with difficulties. Not only is such information limited, but it also is possible that radiation from the nearby Sun has modified the optical properties of mineral grains on Mercury's surface and so rendered conventional interpretations incorrect.

ORIGIN AND EVOLUTION

Mercury's formation. Scientists once thought that Mercury's richness in iron compared with the other terrestrial planets' could be explained by its accretion from objects made up of materials derived from the extremely hot inner region of the solar nebula, where only substances with high freezing temperatures could solidify. The more volatile elements and compounds would not have condensed so close to the Sun. Modern theories of the formation of the solar system, however, discount the possibility that an orderly process of accretion led to progressive detailed differences in planetary chemistry with distance from the Sun. Rather, the components of the bodies that accreted into Mercury likely were derived from a wide part of the inner solar system. Indeed, Mercury itself may have formed anywhere from the asteroid belt inward; subsequent gravitational interactions among the many growing protoplanets could have moved Mercury around.

Speculation about Mercury's high density

Some planetary scientists have suggested that, during Mercury's early epochs, after it had already differentiated (chemically separated) into a less-dense crust and mantle of silicate rocks and a denser iron-rich core, a giant collision stripped away much of the planet's outer layers, leaving a body dominated by its core. This event would have been similar to the collision of a Mars-sized object with Earth that is thought to have formed the Moon (see below *The Moon: Origin and evolution*).

Nevertheless, such violent, disorderly planetary beginnings would not necessarily have placed the inherently densest planet closest to the Sun. Other processes may have been primarily responsible for Mercury's high density. Perhaps the materials that eventually formed Mercury experienced a preferential sorting of heavier metallic particles from lighter silicate ones because of aerodynamic drag by the gaseous solar nebula. Perhaps, because of the planet's nearness to the hot early Sun, its silicates were preferentially vaporized and lost. Each of these scenarios predicts different bulk chemistries for Mercury. In addition, infalling asteroids, meteoroids, and comets and implantation of solar wind particles have been augmenting or modifying the surface and near-surface materials on Mercury for billions of years. Because these materials are the ones most readily analyzed by telescopes and spacecraft, the task of extrapolating backward in time to an understanding of ancient Mercury, and the processes that subsequently shaped it, is formidable.

Later development. Planetary scientists continue to puzzle over the ages of the major geologic and geophysical events that took place on Mercury after its formation. On the one hand, it is tempting to model the planet's history

after that of the Moon, whose chronology has been accurately dated from the rocks returned by the U.S. Apollo manned landings and Soviet Luna robotic missions. By analogy, Mercury would have had a similar history, but one in which the planet cooled off and became geologically inactive shortly after the Caloris impact rather than experiencing persistent volcanism for hundreds of millions of years, as did the Moon. On the presumption that Mercury's craters were produced by the same populations of remnant planetary building blocks (planetesimals), asteroids, and comets that struck the Moon, most of the craters would have formed before and during an especially intense period of bombardment in the inner solar system, which on the Moon is well documented to have ended about 3.8 billion years ago. Caloris presumably would have formed about that time, representing the final chapter in Mercury's geologic history, apart from occasional cratering.

Indications of internal activity

On the other hand, there are many indications that Mercury is very much geologically alive even today. Its dipolar field seems to require a core that is still at least partially molten in order to sustain the magnetohydrodynamic dynamo. Indeed, recent radar measurements of Mercury's spin state have been interpreted as proving that at least the outer core is still molten. In addition, as suggested above, Mercury's scarps show evidence that the planet may not have completed its cooling and shrinking.

There are several approaches to resolving this apparent contradiction between a planet that died geologically before the Moon did and one that is still alive. One hypothesis is that most of Mercury's craters are younger than those on the Moon, having been formed by impacts from so-called vulcanoids—the name bestowed on a hypothetical remnant population of asteroid-sized objects orbiting the Sun inside Mercury's orbit—that would have cratered Mercury over the planet's age. In this case Caloris, the lobate scarps, and other features would be much younger than 3.8 billion years, and Mercury could be viewed as a planet whose surface has only recently become inactive and whose warm interior is still cooling down. No vulcanoids have yet been discovered, however, despite a number of searches for them.

A more likely solution is that the outer shell of Mercury's iron core remains molten because of contamination, for instance, with a small proportion of sulfur, which would lower the melting point of the metal, and of radioactive potassium, which would augment production of heat. Also, the planet's interior may have cooled more slowly than previously calculated as a result of restricted heat transfer. Perhaps the contraction of the planet's crust, so evident about the time of formation of Caloris, pinched off the volcanic vents that had yielded such prolific volcanism earlier in Mercury's history. In this scenario, despite present-day Mercury's lingering internal warmth and churnings, surface activity ceased long ago, with the possible exception of a few thrust faults as the planet continues slowly to contract. (C.R.C.)

Venus

Venus, designated ♀ in astronomy, is the second planet from the Sun and sixth in the solar system in size and mass. No planet approaches closer to Earth than Venus; at its nearest it is the closest large body to Earth other than the Moon. Because Venus's orbit is nearer the Sun than Earth's, the planet is always roughly in the same direction in the sky as the Sun and can be seen only in the hours near sunrise or sunset. When it is visible, it is the most brilliant planet in the sky.

Venus was one of the five planets—along with Mercury, Mars, Jupiter, and Saturn—known in ancient times, and its motions were observed and studied for centuries prior to the invention of advanced astronomical instruments. Its

Association with goddess Ishtar

appearances were recorded by the Babylonians, who equated it with the goddess Ishtar, about 3000 BC, and it also is mentioned prominently in the astronomical records of other ancient civilizations, including those of China, Central America, Egypt, and Greece. Its modern name comes from the Roman goddess of love and beauty (the Greek equivalent being Aphrodite), perhaps because of the planet's luminous jewel-like appearance.

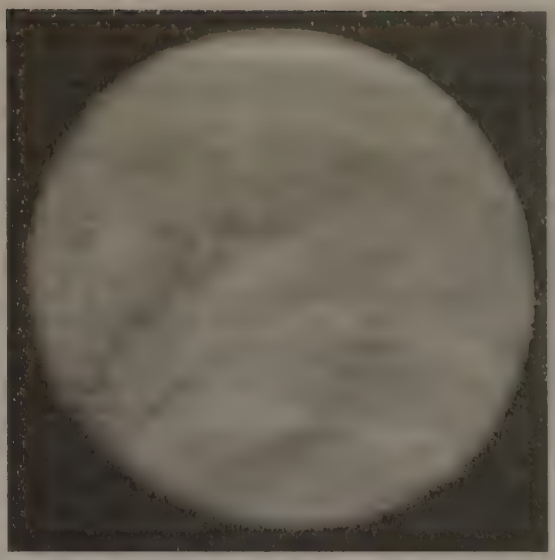


Figure 16: Venus imaged in ultraviolet light by the Pioneer Venus Orbiter spacecraft, Feb. 26, 1979. Although Venus's cloud cover is nearly featureless in visible light, ultraviolet imaging reveals distinctive markings, including global-scale V-shaped bands that open to the west (left).
NASA/JPL

Venus has been called Earth's twin because of the similarities in their masses, sizes, and densities and their similar relative locations in the solar system. Early telescopic observations of the planet revealed a perpetual veil of clouds, suggestive of a substantial atmosphere and leading to popular speculation that Venus was a warm, wet world, perhaps similar to Earth during its prehistoric age of swampy Carboniferous forests and abundant life. Scientists now know, however, that Venus and Earth have evolved surface conditions that could hardly be more different. Venus is extremely hot, dry, and in other ways so forbidding that it is improbable that life as it is understood on Earth could have developed there.

Outdated concept as "swamp world"

BASIC ASTRONOMICAL DATA

Viewed through a telescope, Venus presents a brilliant yellow-white, essentially featureless face to the observer. Its obscured appearance results from the surface of the planet being hidden from sight by a permanent cover of clouds. Features in the clouds are difficult to see in visible light. When observed in ultraviolet light, the clouds exhibit distinctive dark markings, with complex swirling patterns near the equator and global-scale bright and dark bands that are V-shaped and open toward the west. Because of the clouds, little was known about Venus's surface, atmosphere, and evolution before the early 1960s, when the first radar observations were undertaken and spacecraft made the first flybys of the planet.

Venus orbits the Sun at a mean distance of 108 million kilometres, which is about 0.7 times Earth's distance from the Sun. It has the least eccentric orbit of any planet, with a deviation from a perfect circle of only about 1 part in 150. Consequently, its distances at perihelion and aphelion (*i.e.*, when it is nearest and farthest from the Sun, respectively) vary little from the mean distance. The period of its orbit—that is, the length of the Venusian year—is 224.7 Earth days. As Venus and Earth revolve around the Sun, the distance between them varies from a minimum of about 42 million kilometres to a maximum of about 257 million kilometres. Because Venus's orbit lies within Earth's, the planet exhibits phases like those of the Moon when viewed from Earth.

Retrograde rotation

The rotation of Venus on its axis is unusual in both its direction and its speed. The Sun and most of the planets in the solar system rotate in a counterclockwise direction when viewed from above their north poles. Venus, however, rotates in the opposite, or retrograde, direction. Were it

not for the planet's clouds, an observer on Venus's surface would see the Sun rise in the west and set in the east.

Long sidereal day

Venus spins very slowly, taking about 243 Earth days to complete one rotation with respect to the stars—the length of its sidereal day. Venus's spin and orbital periods are very nearly synchronized with Earth's orbit such that, when the two planets are at their closest, Venus presents almost the same face toward Earth. The reasons for this are complex and have to do with the gravitational interactions of Venus, Earth, and the Sun, as well as the effects of Venus's massive rotating atmosphere. Because Venus's spin axis is tilted only about 3° toward the plane of its orbit, the planet does not have appreciable seasons.

Venus's mean radius is 6,051.8 kilometres, or about 95 percent of Earth's at the Equator, while its mass is 4.87×10^{24} kilograms, or 81.5 percent that of Earth. The similarities to Earth in size and mass produce a similarity in density—5.25 grams per cubic centimetre for Venus, compared with 5.52 for Earth. They also result in a comparable surface gravity—humans on Venus would possess about 90 percent of their weight on Earth. Venus is more nearly spherical than most planets. A planet's rotation generally causes a bulging at the equator and a slight flattening at the poles, but Venus's very slow spin allows it to maintain its highly spherical shape. For additional orbital and physical data, see Table 3.

Table 3: Planetary Data for Venus

Mean distance from Sun	108,200,000 km (0.72 AU)
Eccentricity of orbit	0.007
Inclination of orbit to ecliptic	3.4°
Venusian year (sidereal period of revolution)	224.7 Earth days
Rotation period (Venusian sidereal day)	243 Earth days (retrograde)
Mean synodic period	584 Earth days
Mean orbital velocity	35 km/s
Inclination of equator to orbit	117°
Mass	4.87×10^{24} kg
Radius	6,051.8 km
Mean density	5.25 g/cm³
Mean surface gravity	860 cm/s²
Atmospheric composition	96% carbon dioxide, 3.5% molecular nitrogen, 0.02% water, trace quantities of carbon monoxide, molecular oxygen, sulfur dioxide, hydrogen chloride, and other gases
Surface atmospheric pressure	95 bars
Mean surface temperature	737 K (867° F, 464° C)
Mean visible cloud temperature	about 230 K (−46° F, −43° C)
Number of known moons	none

THE ATMOSPHERE

Venus has the most massive atmosphere of the terrestrial planets, which includes Mercury, Earth, and Mars. Its gaseous envelope is composed of more than 96 percent carbon dioxide and 3.5 percent molecular nitrogen. Trace amounts of other gases are present, including carbon monoxide, sulfur dioxide, water vapour, argon, and helium. The atmospheric pressure at the planet's surface varies with surface elevation; at the elevation of the planet's mean radius it is about 95 bars, or 95 times the atmospheric pressure at Earth's surface. This is the same pressure found at a depth of about 1 kilometre in Earth's oceans.

Venus's upper atmosphere extends from the fringes of space down to about 100 kilometres above the surface. There the temperature varies considerably, reaching a maximum of about 300–310 kelvins (K; 80–98° F, 27–37° C) in the daytime and dropping to a minimum of 100–130 K (−280 to −226° F, −173 to −143° C) at night. In the middle atmosphere the temperature increases smoothly with decreasing altitude, from about 173 K (−148° F, −100° C) at 100 kilometres above the surface to roughly 263 K (14° F, −10° C) at the top of the continuous cloud deck, which lies at an altitude of more than 60 kilometres. Below the cloud tops the temperature continues to increase sharply through the lower atmosphere, or troposphere, reaching 737 K (867° F, 464° C) at the surface at the planet's mean radius. This temperature is higher than the melting point of lead.

Temperature variation with altitude

The clouds that enshroud Venus are enormously thick. The main cloud deck rises from about 48 kilometres in altitude to 68 kilometres. In addition, thin hazes exist above and below the main clouds, extending as low as 32 kilometres and as high as 90 kilometres above the surface. The upper haze is somewhat thicker near the poles than elsewhere.

The main cloud deck is formed of three layers. All of them are quite tenuous—an observer in even the densest cloud regions would be able to see objects at distances of several kilometres. The opacity of the clouds varies rapidly with space and time, which suggests a high level of meteorologic activity. The clouds are bright and yellowish when viewed from above, reflecting roughly 85 percent of the sunlight striking them. The material responsible for the yellowish colour has not been confidently identified.

The microscopic particles that make up the Venusian clouds consist of liquid droplets and perhaps also solid crystals. The dominant material is highly concentrated sulfuric acid. Other materials that may exist there include solid sulfur, nitrosylsulfuric acid, and phosphoric acid. Cloud particles range in size from less than 0.5 micrometre (0.00002 inch) in the hazes to a few micrometres in the densest layers.

Sulfuric acid clouds

The circulation of Venus's atmosphere is quite remarkable and is unique among the planets. Although the planet rotates only three times in two Earth years, the cloud features in the atmosphere circle Venus completely in about four days. The wind at the cloud tops blows from east to west at a velocity of about 100 metres per second (360 kilometres per hour). This enormous velocity decreases markedly with decreasing height such that winds at the planet's surface are quite sluggish—typically no more than 1 metre per second (less than 4 kilometres per hour). Much of the detailed nature of the westward flow above the cloud tops can be attributed to tidal motions induced by solar heating. Nevertheless, the fundamental cause of this "superrotation" of Venus's dense atmosphere is unknown, and it remains one of the more intriguing mysteries in planetary science.

Most information about wind directions at the planet's surface comes from observations of wind-blown materials. Despite low surface-wind velocities, the great density of Venus's atmosphere enables these winds to move loose fine-grained materials, producing surface features that have been seen in radar images. Some features resemble sand dunes, while others are "wind streaks" produced by preferential deposition or erosion downwind from topographic features. The directions assumed by the wind-related features suggest that in both hemispheres the surface winds blow predominantly toward the equator. This pattern is consistent with the idea that simple hemispheric-scale circulation systems called Hadley cells exist in the Venusian atmosphere. According to this model, atmospheric gases rise upward as they are heated by solar energy at the planet's equator, flow at high altitude toward the poles, sink to the surface as they cool at higher latitudes, and flow toward the equator along the planet's surface until they warm and rise again.

A major consequence of Venus's massive atmosphere is that it produces an enormous greenhouse effect, which intensely heats the planet's surface. Because of its bright continuous cloud cover, Venus actually absorbs less of the Sun's light than does Earth. Nevertheless, the sunlight that does penetrate the clouds is absorbed in the lower atmosphere and at the surface. The surface and the gases of the lower atmosphere, which are heated by the absorbed light, reradiate this energy at infrared wavelengths. On Earth most reradiated infrared radiation escapes back into space, which allows Earth to maintain a reasonably cool surface temperature. On Venus the dense carbon dioxide atmosphere and the thick cloud layers trap much of the infrared radiation. The trapped radiation heats the lower atmosphere further, ultimately raising the surface temperature by hundreds of degrees. Study of the Venusian greenhouse effect has led to an improved understanding of the more subtle but very important influence of greenhouse gases in Earth's atmosphere and a greater appreciation of the effects of energy use and of other human activities on Earth's energy balance. (See below *Earth: The atmosphere and hydrosphere: The atmosphere*.)

Greenhouse effect

Above the main body of the Venusian atmosphere lies the

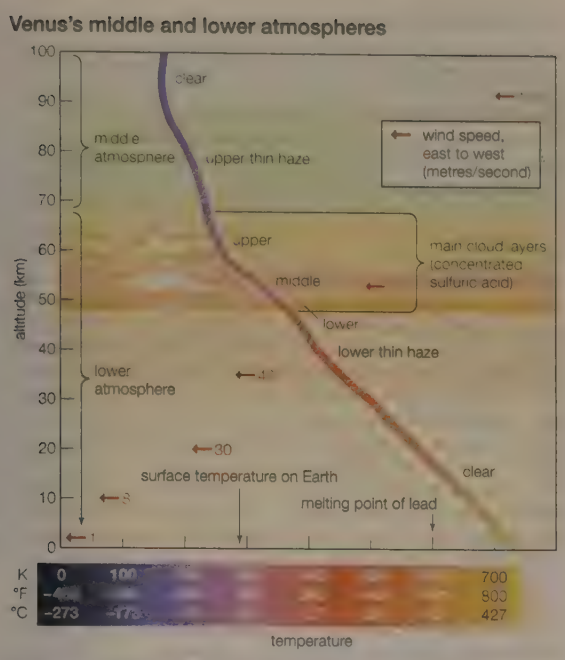


Figure 17: Profile of Venus's middle and lower atmospheres as derived from measurements made by the Pioneer Venus mission's atmospheric probes and other spacecraft.

ionosphere. As its name implies, the ionosphere is composed of ions, or charged particles, produced both by absorption of ultraviolet solar radiation and by the impact of the solar wind—the flow of charged particles streaming outward from the Sun—on the upper atmosphere. The primary ions in the Venusian ionosphere are forms of oxygen (O^+ and O_2^+) and carbon dioxide (CO_2^+).

INTERACTION WITH THE SOLAR WIND

Absence of magnetic field

Unlike most planets, including Earth, Venus does not exhibit an intrinsic magnetic field. Sensitive measurements by orbiting spacecraft have shown that any dipole field originating from within Venus must be no more than 1/8,000 that of Earth's. The lack of a magnetic field may be related in part to the planet's slow rotation because, according to the dynamo theory that explains the origin of planetary magnetic fields, rotation helps to drive the fluid motions within the planet's interior that produce the field. It is also possible that Venus may lack a magnetic field because its core is fluid but does not circulate or simply because the core is solid and hence is incapable of supporting a dynamo.

As the solar wind bombards a planet at supersonic speeds, it generally forms a bow shock on the planet's sunward side—that is, a standing wave of plasma that slows down, heats, and deflects the flow around the planet. For planets such as Jupiter and Earth, the bow shock lies at a considerable distance from the surface, held off by the planet's magnetic field. Because Venus lacks a detectable field, however, its bow shock lies just a few thousand kilometres above the surface, held off only by the planet's ionosphere. This closeness of the bow shock to the surface leads to particularly intense interactions between the solar wind and Venus's atmosphere. In fact, the top of the ionosphere, known as the ionopause, lies at a much lower altitude on the dayside of Venus than on the nightside, because of the pressure exerted by the solar wind. The density of the ionosphere is also far greater on the dayside of the planet than on the nightside.

Mechanism for water loss

Venus's interaction with the solar wind results in a gradual, continuous loss to space of hydrogen and oxygen from the planet's upper ionosphere. This process is equivalent to a gradual loss of water from the planet. Over the course of Venus's history, the total amount of water lost via this mechanism could have been as much as a few percent of a world ocean the size of Earth's.

CHARACTER OF THE SURFACE

The high atmospheric pressure, the low wind velocities, and, in particular, the extremely high temperatures create a surface environment on Venus that is markedly different from any other in the solar system. A series of landings by robotic Soviet spacecraft in the 1970s and early '80s provided detailed data on surface composition and appearance. Views of the Venusian landscape, typified in colour images obtained by the Soviet Union's Venera 13 lander in 1982 (see Figure 18), show rocky plains that stretch toward the horizon. Despite the heavy cloud cover, the surface is well illuminated by the yellow-orange light that filters through the clouds.

Yellow-orange surface light

The most striking characteristic of the surface at the Venera 13 site and most other Venera landing sites is the flat, slabby, layered nature of the rocks. Both volcanic and sedimentary rocks on Earth can develop such an appearance under appropriate conditions, but the reason that the Venusian rocks have done so is not known with certainty. Also present among the rocks is a darker, fine-grained soil. The grain size of the soil is unknown, but some of it was fine enough to be lifted briefly into the atmosphere by the touchdown of the Venera lander, which suggests that some grains are no more than a few tens of micrometres in diameter. Scattered throughout the soil and atop the rocks are pebble-size particles that could be either small rocks or clods of soil.

The general surface appearance at the Venera landing sites is probably common on Venus, but it is likely not representative of all locations on the planet. Radar data from the U.S. Magellan spacecraft, which studied Venus from orbit in the early 1990s, provided global information about

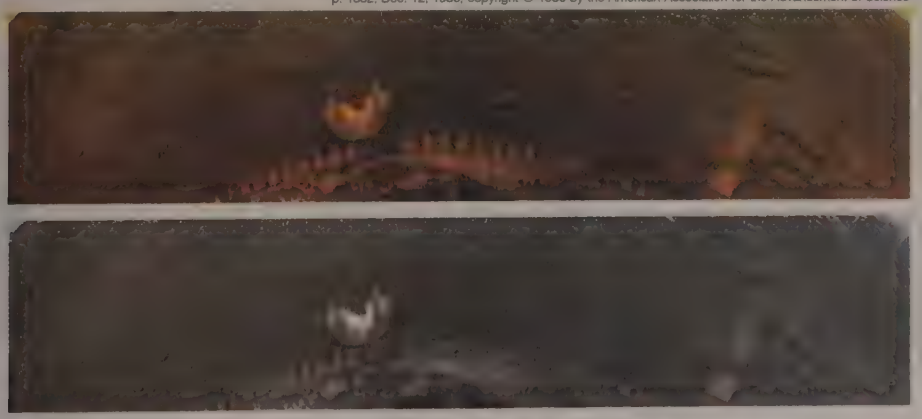


Figure 18: Panoramic 170° view of the surface of Venus—in natural light (top) and colour-corrected to appear as it would in white light (bottom)—obtained by the Venera 13 lander on March 1, 1982. Flat rock slabs and soil extend to a horizon that can just barely be seen in the extreme upper corners of the images, which appear curved because of the lander's camera scanning pattern. Parts of the spacecraft are visible in the lower half of the images.

the roughness of the Venusian surface at scales of metres to tens of metres. Although much of the planet is indeed covered by lowland plains that appear smooth to radar, some terrains were found to be very much rougher. These include areas covered by ejecta (the material expelled from impact craters and extending around them), steep slopes associated with tectonic activity, and some lava flows. How such terrains would appear from a lander's perspective is not known, but large boulders and other sorts of angular blocks presumably would be more common than at the Venera sites.

Surface composition. A number of the Soviet landers carried instruments to analyze the chemical composition of the surface materials of Venus. Because only the relative proportions of a few elements were measured, no definitive information exists concerning the rock types or minerals present. Two techniques were used to measure the abundances of various elements. Gamma-ray spectrometers, which were carried on Veneras 8, 9, and 10 and the landers of the Soviet Vega 1 and 2 missions, measured the concentrations of naturally radioactive isotopes of the elements uranium, potassium, and thorium. X-ray fluorescence instruments, carried on Veneras 13 and 14 and Vega 2, measured the concentrations of a number of major elements.

The Venera 8 site gave indications that the rock composition may be similar to that of granite or other igneous rocks that compose Earth's continents. This inference, however, was based only on rather uncertain measurements of the concentrations of a few radioactive elements.

Rock similarities to basalts on Earth

Measurements of radioactive elements at the Venera 9 and 10 and Vega 1 and 2 landing sites suggested that the compositions there resemble those of basalt rocks found on Earth's ocean floors and in some volcanic regions such as Hawaii and Iceland. The Venera 13 and 14 and Vega 2 X-ray instruments measured concentrations of silicon, aluminum, magnesium, iron, calcium, potassium, titanium, manganese, and sulfur. Although some composition differences were seen among the three sites, on the whole the elemental compositions measured by all three landers were similar to those of basalts on Earth.

A surprising result of orbital radar observations is that the highest elevations on Venus exhibit anomalously high radar reflectivity. The best interpretation seems to be that the highest elevations are coated with a thin layer of some semiconducting material. Its composition is unknown, but it could be an iron-containing mineral such as pyrite or magnetite, which formed at cooler, higher elevations from low concentrations of atmospheric iron(II) chloride vapour in the atmosphere.

Surface features. Earth-based observatories and Venus-orbiting spacecraft have provided global-scale information on the nature of the planet's surface. All have used radar systems to penetrate the thick Venusian clouds.

The entire surface of the planet is dry and rocky. Because there is no sea level in the literal sense, elevation is expressed as a planetary radius—*i.e.*, as the distance from the centre of the planet to the surface at a given location. Most of the planet consists of gently rolling plains. Globally, more than 80 percent of the surface deviates less than 1 kilometre from the mean radius. At several locations on the plains are broad, gently sloping topographic depressions, or lowlands, that may reach several thousand kilometres across; they include Atalanta Planitia, Guinevere Planitia, and Lavinia Planitia.

Continent-sized highlands

Two striking features are the continent-sized highland areas, or terrae—Ishtar Terra in the northern hemisphere and Aphrodite Terra along the equator. Ishtar is roughly the size of Australia, while Aphrodite is comparable in area to South America. Ishtar possesses the most spectacular topography on Venus. Much of its interior is a high plateau, called Lakshmi Planum, that resembles in configuration the Plateau of Tibet on Earth. Lakshmi is bounded by mountains on most sides, the largest range being the enormous Maxwell Montes on the east. These mountains soar about 11 kilometres above the mean radius of Venus. The topography of Aphrodite, more complex than that of Ishtar, is characterized by a number of distinct mountain ranges and several deep, narrow troughs. In addition to the two main terrae are several smaller elevated regions, including Alpha Regio, Beta Regio, and Phoebe Regio.

Many of the surface features on Venus can be attributed to tectonic activity—that is, to deformational motions within the crust. These include mountain belts, plains deformation belts, rifts, coronae, and tesserae, which are discussed in turn below.

Mountain belts. Found in the terrae, Venus's mountain belts are in some ways similar to ones on Earth, such as the Himalayas of Asia and the Andes of South America. Among the best examples are those that encircle Lak-

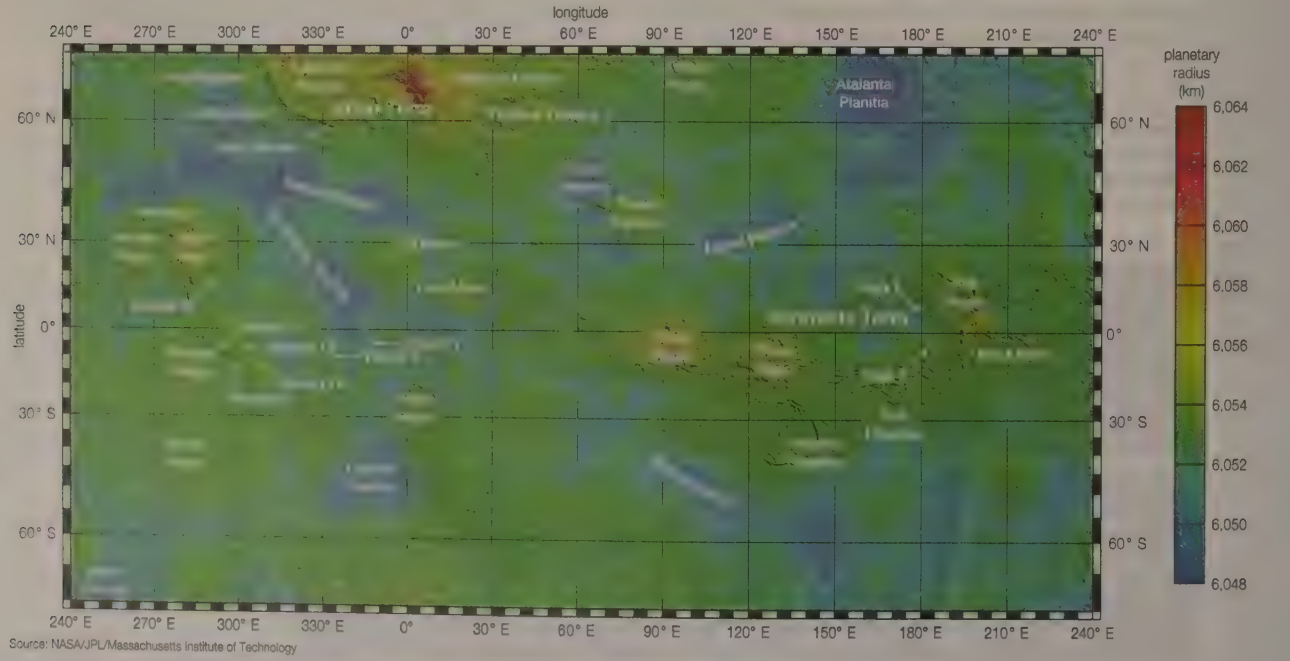


Figure 19: Global topographic map of Venus derived from laser altimetry data gathered by the Magellan spacecraft, which carried out observations from orbit around the planet between 1990 and 1994. This Mercator projection extends to latitudes 70° N and S. Relief is colour-coded according to the key at right, with values expressed as distance from the centre of the planet. Selected major topographic features and spacecraft landing sites are labeled.

shmi Planum, which in addition to Maxwell Montes include Freyja, Akna, and Danu Montes. Maxwell Montes is particularly broad and comparable in size to the Himalayas.

Venusian mountain building

Venus's mountain belts typically consist of parallel ridges and troughs with spacings of 5–10 kilometres. They probably developed when broad bands of the lithosphere were compressed from the sides and became thickened, folding and thrusting surface materials upward. Their formation in some respects thus resembles the building of many mountain ranges on Earth. On the other hand, because of the lack of liquid water or ice on Venus, their appearance differs in major ways from their counterparts on Earth. Without the flow of rivers or glaciers to wear them down, Venusian mountain belts have acquired steep slopes as a result of folding and faulting. The erosional forms common in mountainous regions on Earth are absent.

Plains deformation belts. Although plains deformation belts are similar in some ways to mountain belts, they display less pronounced relief and are found primarily in low-lying areas of the planet, such as Lavinia Planitia and Atalanta Planitia. Like mountain belts, they show strong evidence for parallel folding and faulting and may form primarily by compression, deformation, and uplift of the lithosphere. Within a given lowland, it is common for deformation belts to lie roughly parallel to one another, spaced typically several hundred kilometres apart.

Rifts. Rifts are among the most spectacular tectonic features on Venus. The best-developed rifts are found atop broad, raised areas such as Beta Regio, sometimes radiating outward from their centres like the spokes of a giant wheel. Beta and several other similar regions on Venus appear to be places where large areas of the lithosphere have been forced upward from below, splitting the surface to form great rift valleys. The rifts are composed of innumerable faults, and their floors typically lie 1–2 kilometres below the surrounding terrain. In many ways the rifts on Venus are similar to great rifts elsewhere, such as the East African Rift on Earth or Valles Marineris on Mars; volcanic eruptions, for example, appear to have been associated with all these features. The Venusian rifts differ from Earth and Martian ones, however, in that little erosion has taken place within them, because of the lack of water.

Coronae. Coronae (Latin: "garlands" or "crowns") are landforms that apparently owe their origin to the effects of hot, buoyant blobs of material, known from terrestrial geology as diapirs, that originate deep beneath the surface of Venus. Coronae evolve through several stages. As diapirs first rise through the planet's interior and approach the surface, they can lift the rocks above them, fracturing the surface in a radial pattern. This results in a distinctive starburst of faults and fractures, often lying atop a broad, gently sloping topographic rise. Such features are sometimes called novae, a name given to them when their evolutionary relationship to coronae was less certain. Once a diapir has neared the surface and cooled, it loses its buoyancy. The initially raised crust then can sag under its own weight, developing concentric faults as it does so. The result is a circular-to-oval pattern of faults, fractures, and ridges. Volcanism can occur through all stages of corona formation.

Relationship to novae

Coronae are typically a few hundred kilometres in diameter. Although they may have a raised outer rim, many coronae sag noticeably in their interiors and also outside their rims. Hundreds of coronae are found on Venus, observed at all stages of development. The radially fractured domes of the early stages are comparatively uncommon, while the concentric scars characteristic of mature coronae are among the most numerous large tectonic features on the planet.

Tesserae. Tesserae (Latin: "mosaic tiles") are the most geologically complex regions seen on Venus. Several large elevated regions, such as Alpha Regio, are composed largely of tessera terrain. Such terrain appears extraordinarily rugged and highly deformed in radar images, and in some instances it displays several different trends of parallel ridges and troughs that cut across one another at a wide range of angles. The deformation in tessera terrain can be so complex that sometimes it is difficult to determine what kinds of stresses in the lithosphere were responsible for forming it. In fact, probably no single process can explain all tessera formation. Tesserae typically appear very bright in radar images, which suggests an extremely rough and blocky surface at scales of metres. Some tesserae may be old terrain that has been subjected to more episodes of mountain building and faulting than have the materials around it, each one superimposed on its predecessor to produce the pattern observed.

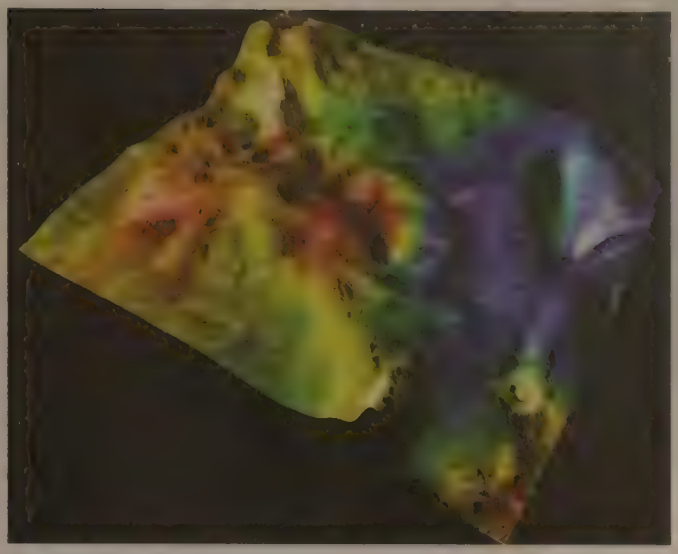


Figure 20: Oblique, vertically exaggerated view of coronae in the Sedna Planitia lowlands of Venus, generated and colour-coded by computer from data collected by the Magellan spacecraft. The rise left of centre is a corona in an early evolutionary stage (sometimes called a nova), characterized by raised crust that is fractured in a radial pattern. The depression at the far right is a corona in a later stage, in which the raised crust has sagged at the centre, with concentric fractures added to the radial ones.
NASA/JPL/Caltech

Volcanic features. Along with intense tectonic activity, Venus has undergone much volcanism. The largest volcanic outpourings are the huge lava fields that cover most of the rolling plains. These are similar in many respects to fields of overlapping lava flows seen on other planets, including Earth, but they are far more extensive. Individual flows are for the most part long and thin, which indicates that the erupting lavas were very fluid and hence were able to flow long distances over gentle slopes. Lavas on Earth and the Moon that flow this readily typically consist of basalts, and so it is probable that basalts are common on the plains of Venus as well.

Lava-carved *canali*

Of the many types of lava-flow features seen on the Venusian plains, none are more remarkable than the long, sinuous *canali*. These meandering channels usually have remarkably constant widths, which can be as much as 3 kilometres. They commonly extend as far as 500 kilometres across the surface; one is 6,800 kilometres long. *Canali* probably were carved by very low-viscosity lavas that erupted at sustained high rates of discharge. Other channel-like volcanic features on Venus include sinuous rilles that may be collapsed lava tubes, and large, complex compound valleys that apparently result from particularly massive outpourings of lava.

Shield volcanoes

In many locations on Venus, volcanic eruptions have built edifices similar to the great volcanoes of Hawaii on Earth or those associated with the Tharsis region on Mars. Sif Mons (Figure 21) is an example of such a volcano; there are more than 100 others distributed widely over the planet. Known as shield volcanoes, they reach heights of several kilometres above the surrounding plains and can be hundreds of kilometres across at their base. They are made up of many individual lava flows piled on one another in a radial pattern. They develop when a source of lava below the surface remains fixed and active at one location long enough to allow the volcanic materials it extrudes to accumulate above it in large quantities. Like those found on the

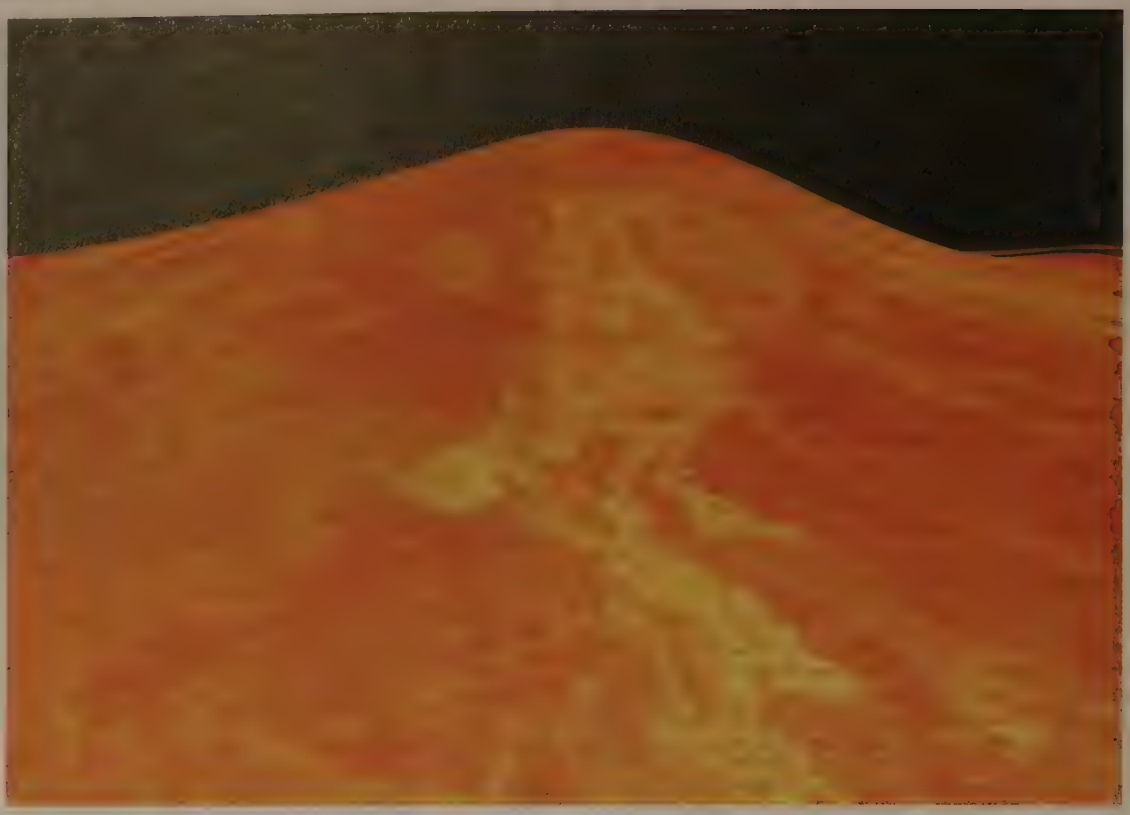


Figure 21: Sif Mons, a shield volcano on Venus, in a low-angle, vertically exaggerated, computer-generated view derived from radar and altimetry data from the Magellan spacecraft. The volcano is about 2 kilometres high and has a base 300 kilometres in diameter. Added colour in the image is based on photos recorded by Soviet Venera landers.
NASA/JPL

rolling plains, the flows constituting the shield volcanoes are generally very long and thin and are probably composed of basalt.

When a subsurface source of lava is drained of its contents, the ground above it may collapse, forming a depression called a caldera. Many volcanic calderas are observed on Venus, both atop shield volcanoes and on the widespread lava plains. They are often roughly circular in shape and overall are similar to calderas observed on Earth and Mars. The summit region of Sif Mons, for example, exhibits a caldera-like feature 40–50 kilometres in diameter.

Along with the extensive lava plains and the massive shield volcanoes are many smaller volcanic landforms. Enormous numbers of small volcanic cones are distributed throughout the plains. Particularly unusual in appearance are so-called pancake domes, which are typically a few tens of kilometres in diameter and about 1 kilometre high and are remarkably circular in shape. Flat-topped and steep-sided, they appear to have formed when a mass of unusually thick lava was extruded from a central vent and spread outward for a short distance before solidifying. The composition of this lava is unknown, but—given the knowledge of lavas on Earth—it is likely to be much richer in silica than the basalt thought to predominate elsewhere on the planet.

Pancake domes

NASA

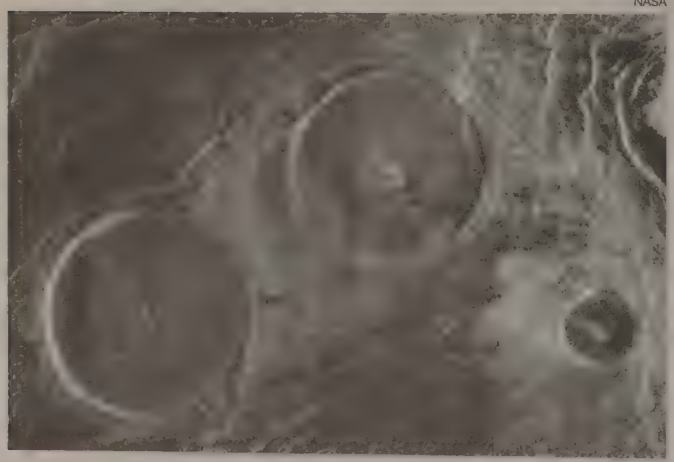


Figure 22: Volcanic pancake domes in the elevated region Eistla Regio on Venus, in a radar image produced from Magellan spacecraft data. The two larger domes, each about 65 kilometres across, have broad flat tops less than 1 kilometre high. All three apparently were formed from unusually thick lava that oozed to the surface and spread in all directions.

Volcanic edifices are not uniformly distributed on Venus. Although they are common everywhere, they are particularly concentrated in the Beta-Atla-Themis region, between longitudes 180° and 300° E. This concentration may be the consequence of a broad active upwelling of the Venusian mantle in this area, which has led to enhanced heat flow and formation of magma reservoirs.

Impact craters. The Venusian surface has been altered by objects from outside the planet as well as by forces from within. Impact craters dot the landscape, created by meteorites that passed through the atmosphere and struck the surface. Nearly all solid bodies in the solar system bear the scars of meteoritic impacts, with small craters typically being more common than large ones. This general tendency is encountered on Venus as well—craters a few hundred kilometres across are present but rare, while craters tens of kilometres in diameter and smaller are common. Venus has an interesting limitation, however, in that craters smaller than about 1.5–2 kilometres in diameter are not found. Their absence is attributable to the planet's dense atmosphere, which causes intense frictional heating and strong aerodynamic forces as meteorites plunge through it at high velocities. The larger meteorites reach the surface intact, but the smaller ones are fragmented in the atmosphere. In fact, craters several kilometres in size—*i.e.*, near the minimum size observed—tend not to be circular. Instead they have complex shapes, often with several

Effect of dense atmosphere

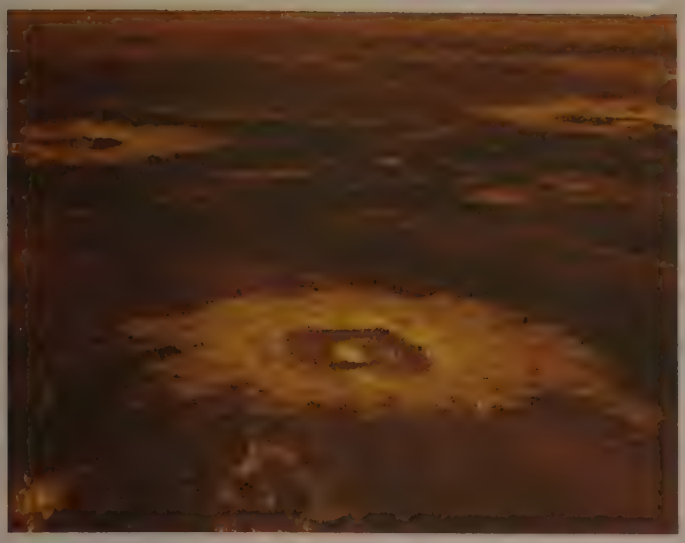


Figure 23: A trio of impact craters in Lavinia Planitia, a lowland plain in the southern hemisphere of Venus, shown in a computer-generated image created from Magellan spacecraft radar data. The craters range between about 40 and 60 kilometres across and are of average size for the planet. The surrounding lobed ejecta blankets stand out in the radar image as bright (and hence comparatively rough) terrain. Added colour in the image is based on surface images taken by Soviet Venera landers.
NASA/JPL

irregular pits rather than a single depression, which suggests that the impacting body broke up into a number of pieces that struck the surface individually. Radar images also show diffuse dark and bright "smudges" that may have resulted from the explosions of small meteorites above the surface.

The large craters seen on Venus are different in some respects from those observed on other planets. Most impact craters, on Venus and elsewhere, show ejecta around them.

Lobed ejecta

Venusian ejecta is unusual, however, in that its outer border commonly shows a lobed or flower-petal pattern, which suggests that much of it poured outward in a ground-hugging flow rather than arcing high above the ground and falling back to the surface. This behaviour was probably produced by dense atmospheric gases that became entrained in the flow and resulted in a turbulent cloud of gas and ejecta. Another peculiarity of large Venusian craters is the sinuous flows that have emerged from the ejecta, spreading outward from it just as lava flows would. These flows are apparently composed of rock that was melted by the high pressures and temperatures reached during the impact. The prevalence of these flow features on Venus must be due in large part to the high surface temperature—rocks are closer to their melting temperature when craters form, which allows more melt to be produced than on other planets. For the same reason, the molten rock will remain fluid longer, which allows it to flow for significant distances.

Unique parabola-shaped deposits

Perhaps the strangest property of Venusian craters is one associated with some of the youngest. In addition to the normal ejecta, these craters are partially surrounded by huge parabola-shaped regions of dark material, a feature not found elsewhere in the solar system. In every case, the parabola opens to the west, and the crater is nestled within it, toward its eastern extremity. In radar images the dark materials tend to be smooth at small scales; it is likely that these parabolas are composed of deposits of fine-grained ejecta that was thrown upward during the impact. Apparently the material rose above the Venusian atmosphere, fell back, and was picked up by the high-speed westward-blowing winds that encircle the planet. It was then carried far downwind from the impact site, eventually descending to the surface to form a parabola-shaped pattern.

For planets and moons that have impact craters, crater populations are an important source of information about the ages of the surfaces on which they lie. The concept is simple in principle—on a given body older surfaces have more craters than do younger ones. Determining an absolute age in years is difficult, however, and requires knowledge about the rate of crater formation, which usually must be inferred indirectly. The absolute ages of materials on the surface of Venus are not known, but the overall density of craters on Venus is lower than on many other bodies in the solar system. Estimates vary, but the average age of materials on Venus is almost certainly less than one billion years and may be substantially less.

The spatial distribution of craters on Venus is essentially random. If craters were clustered in distinct regions, scientists could infer that a wide range of surface ages was represented over the planet. With a near-random global crater distribution, however, they are led instead to the conclusion that essentially the entire planet has been geologically resurfaced in the last billion years or less and that much of the resurfacing took place in a comparatively brief time.

INTERIOR STRUCTURE AND GEOLOGIC EVOLUTION

Much less is known about the interior of Venus than about its surface and atmosphere. Nevertheless, because the planet is much like Earth in overall size and density and because it presumably accreted from similar materials, scientists expect that it evolved at least a crudely similar internal state. Therefore, it probably has a core of metal, a mantle of dense rock, and a crust of less-dense rock. The core, like that of Earth, is probably composed primarily of iron and nickel, although Venus's somewhat lower density may indicate that its core also contains some other, less-dense material such as sulfur. Calculations of Venus's internal structure suggest that the outer boundary of the core lies a little more than 3,000 kilometres from the centre of the planet.

Above the core and below the crust lies Venus's mantle, making up the bulk of the planet's volume. Despite the

NASA/JPL

Figure 24: Adivar Crater on Venus, in a radar image from the Magellan spacecraft. About 30 kilometres in diameter, the impact scar is surrounded by a characteristic ejecta blanket. Unusual, however, is the much larger region affected by the impact, which includes wind-carried materials, bright in radar images and distributed mostly to the west (left) of the crater, and a surrounding radar-dark, westward-opening parabolic border.

high surface temperatures, temperatures within the mantle are likely similar to those in Earth's mantle. Even though a planetary mantle is composed of solid rock, the material there can slowly creep or flow, just as glacial ice does, allowing sweeping convective motions to take place. Convection is a great equalizer of the temperatures of planetary interiors. Similar to heat production within Earth, heat within Venus is thought to be generated by the decay of natural radioactive materials. This heat is transported to the surface by convection. If temperatures deep within Venus were substantially higher than those within Earth, the viscosity of the rocks in the mantle would drop sharply, speeding convection and removing the heat more rapidly. Therefore, the deep interiors of Venus and Earth are not expected to differ dramatically in temperature.

Internal heat generation

As noted above, the composition of the Venusian crust is thought to be dominated by basalt. Gravity data suggest that the thickness of the crust is fairly uniform over much of the planet, with typical values of perhaps 20–50 kilometres. Possible exceptions are the tessera highlands, where the crust may be significantly thicker.

Convective motions in a planet's mantle can cause materials near the surface to experience stress, and motions in the Venusian mantle may be largely responsible for the tectonic deformation observed in radar images. Raised topography, such as Beta Regio, could lie above regions of mantle upwelling, whereas lowered topography, such as Lavinia Planitia, could lie above regions of mantle downwelling.

Despite the many overall similarities between Venus and Earth, the geologic evolution of the two planets has been strikingly different. Evidence suggests that the process of plate tectonics (see below *Earth: The outer shell*) does not now operate on Venus. Although deformation of the lithosphere does indeed seem to be driven by mantle motions, lithospheric plates do not move mainly horizontally relative to each other, as they do on Earth. Instead, motions are mostly vertical, with the lithosphere warping up and down in response to the underlying convective motions. Volcanism, coronae, and rifts tend to be concentrated in regions of upwelling, while plains deformation belts are concentrated in regions of downwelling. The formation of rugged uplands such as Aphrodite and Ishtar is not as well understood, but the mechanism probably involves some kind of local crustal thickening in response to mantle motions.

The lack of plate tectonics on Venus may in part be due to the planet's high surface temperature, which makes the upper rigid layer of the planet—the lithosphere—more buoyant and hence more resistant to subduction than Earth's lithosphere, other factors being equal. Interestingly, there is evidence that the Venusian lithosphere may be thicker than Earth's and that it has thickened with time. A gradual, long-term thickening of Venus's lithosphere in fact could be related to the curious conclusion drawn from Venus's cratering record (see above *Impact craters*)—that most of the planet underwent a brief but intense period of geologic resurfacing less than a billion years ago. One possible explanation is that Venus may experience episodic global overturns of its mantle, in which an initially thin lithosphere slowly thickens until it founders on a near-global scale, triggering a brief, massive geologic resurfacing event. How many times this may have occurred during the planet's history and when it may happen again are unknown.

Geologically recent resurfacing episode

OBSERVATIONS FROM EARTH

Since the Italian scientist Galileo's discovery of Venus's phases in 1610, the planet has been studied in detail, using Earth-based telescopes, radar, and other instruments. Important early telescopic observations were conducted in the 1700s during the planet's solar transits. In a solar transit an object passes directly between the Sun and Earth and is silhouetted briefly against the Sun's disk. Transits of Venus are rare events, occurring in pairs eight years apart with more than a century between pairs. They were extremely important events to 18th-century astronomy, because they provided the most accurate method known at that time for determining the distance between Earth and the Sun. (This distance, known as the astronomical unit, is a fundamental unit of astronomy.)

Observations of the 1761 transit were only partially successful but did result in the first suggestion, by the Russian scientist Mikhail V. Lomonosov, that Venus has an atmosphere. The second transit of the pair, in 1769, was observed with somewhat greater success. Transits must be viewed from many points on Earth to yield accurate distances, and the transits of 1761 and, particularly, 1769 prompted the launching of many scientific expeditions to remote parts of the globe. Among these was the first of the three voyages of exploration by the British naval officer James Cook, who, with scientists from the Royal Society, observed the 1769 transit from Tahiti. The transit observations of the 1700s not only gave an improved value for the astronomical unit but also provided the impetus for many unrelated but important discoveries concerning Earth's geography.

Transit of 1769

By the time the subsequent pair of transits occurred, in 1874 and 1882, the nascent field of celestial photography had advanced enough to allow scientists to record on glass plates what they saw through their telescopes. No transits took place in the 20th century; the first of the next pair was widely observed and imaged in 2004.

In the modern era Venus has also been observed at wavelengths outside the visible spectrum. The cloud features were discovered with certainty in 1927–28 in ultraviolet photographs. The first studies of the infrared spectrum of Venus, in 1932, showed that its atmosphere is composed primarily of carbon dioxide. Subsequent infrared observations revealed further details about the composition of both the atmosphere and the clouds. Observations in the microwave portion of the spectrum, beginning in earnest in the late 1950s and early '60s, provided the first evidence of the extremely high surface temperatures on the planet and prompted the study of the greenhouse effect as a means of producing these temperatures.

After finding that Venus is wrapped in clouds, astronomers turned to other techniques to study its surface. Foremost among these has been radar. If equipped with an appropriate transmitter, a large radio telescope can be used as a radar system to bounce a radio signal off a planet and detect its return. Because radio wavelengths penetrate the Venusian atmosphere, the technique is an effective means of probing the planet's surface.

Use of radar

Earth-based radar observations have been conducted primarily from Arecibo Observatory in the mountains of Puerto Rico, the Goldstone tracking station complex in the desert of southern California, and Haystack Observatory in Massachusetts. The first successful radar observations of Venus took place at Goldstone and Haystack in 1961 and revealed the planet's slow rotation. Subsequent observations determined the rotation properties more precisely and began to unveil some of the major features on the planet's surface. The first features to be observed were dubbed Alpha, Beta, and Maxwell, the last after James Clerk Maxwell, the British physicist who first derived some of the basic equations that describe the propagation of electromagnetic radiation. These three features are among the brightest on the planet in radar images, and their names have been preserved to the present as Alpha Regio, Beta Regio, and Maxwell Montes.

By the mid-1980s Earth-based radar technology had advanced such that images from Arecibo were revealing features as small as a few kilometres in size. Nevertheless, because Venus always presents nearly the same face toward Earth when the planets are at their closest, much of the surface went virtually unobserved from Earth.

SPACECRAFT EXPLORATION

The greatest advances in the study of Venus were achieved through the use of robotic spacecraft. The first spacecraft to reach the vicinity of another planet and return data was the U.S. Mariner 2 in its flyby of Venus in 1962. Since then Venus has been the target of more than 20 spacecraft missions.

Successful early Venus missions undertaken by the United States involved Mariner 2, Mariner 5 (1967), and Mariner 10 (1974). Each spacecraft made a single close

flyby, providing successively improved scientific data. After visiting Venus, Mariner 10 went on to a successful series of flybys of Mercury. In 1978 the United States launched the Pioneer Venus mission, comprising two complementary spacecraft. The Orbiter went into orbit around the planet, while the Multiprobe released four entry probes—one large probe and three smaller ones—that were targeted to widely separated points in the Venusian atmosphere to collect data on atmospheric structure and composition. The Orbiter carried 17 scientific instruments, most of them focused on study of the planet's atmosphere, ionosphere, and interaction with the solar wind. Its radar altimeter provided the first high-quality map of Venus's surface topography.

Venera missions

Venus was also a major target of the Soviet Union's planetary exploration program during the 1960s, '70s, and '80s, which achieved several spectacular successes. After an early sequence of failed missions, in 1967 Soviet scientists launched Venera 4, comprising a flyby spacecraft as well as a probe that entered the planet's atmosphere. Highlights of subsequent missions included the first successful soft landing on another planet (Venera 7 in 1970), the first images returned from the surface of another planet (Venera 9 and 10 landers in 1975), and the first spacecraft placed in orbit around Venus (Venera 9 and 10 orbiters).

In terms of the advances they provided in the global understanding of Venus, the most important Soviet missions were Veneras 15 and 16 in 1983. The twin orbiters carried the first radar systems flown to another planet that were capable of producing high-quality images of the surface. They mapped the northern quarter of Venus with a resolution of 1–2 kilometres, and many types of geologic features now known to exist on the planet were either discovered or first observed in detail in the Venera 15 and 16 data. Late the following year the Soviet Union launched Vegas 1 and 2. These delivered Venera-style landers and dropped off two balloons in the Venusian atmosphere, each of which survived for about two days and transmitted data from their float altitudes in the middle cloud layer. The Vega spacecraft themselves continued on to make successful flybys of Halley's Comet in 1986.

In 1990, on its way to Jupiter, the U.S. Galileo spacecraft flew by Venus. Among its most notable observations were images at near-infrared wavelengths that viewed deep into the atmosphere and showed the highly variable opacity of the main cloud deck.

Magellan observations

The most ambitious mission yet to Venus, the U.S. Magellan spacecraft, was launched in 1989 and the next year entered orbit around the planet, where it conducted observations until late 1994. Magellan carried a radar system capable of producing images with a resolution better than 100 metres. Because the orbit was nearly polar, the spacecraft was able to view essentially all latitudes on the planet. On each orbit the radar system obtained an image strip about 20 kilometres wide and typically more than 16,000 kilometres long, extending nearly from pole to pole. The strips were assembled into mosaics, and high-quality radar images of about 98 percent of the planet were ultimately produced. Magellan also carried a radar altimeter that measured the planet's surface topography as well as some properties of its surface materials. After the main radar objectives of the mission were completed, the spacecraft's orbit was modified slightly so that it passed repeatedly through the upper fringes of the Venusian atmosphere. The resulting drag on the spacecraft gradually removed energy from its orbit, turning an initially elliptical orbit into a low, circular one. This procedure, known as aerobraking, has since been used on other planetary missions to conserve large amounts of fuel by reducing the use of thrusters for orbital reshaping. From its new circular orbit, the Magellan spacecraft was able to make the first detailed map of Venus's gravitational field. (S.W.S.)

Earth

Earth, designated ⊕ in astronomy, is the third planet from the Sun and the fifth in the solar system in size and mass. Its single most outstanding feature is that its near-surface environments are the only places in the universe known to harbour life. Earth's name in English, the international language of astronomy, derives from Old English and Germanic words for *ground* and *earth*. It is the only name for a planet of the solar system that does not come from Greco-Roman mythology.

Earth viewed from space

Viewed from another planet in the solar system, Earth would appear bright and bluish in colour. Easiest to see through a large telescope would be its atmospheric features, chiefly the swirling white cloud patterns of midlatitude and tropical storms, ranged in roughly latitudinal belts around the planet. The polar regions also would appear a brilliant white, because of the clouds above and the snow and ice below. Beneath the changing patterns of clouds would appear the much darker blue-black oceans, interrupted by occasional tawny patches of desert lands. The green landscapes that harbour most human life would not be easily seen from space. Not only do they constitute a modest fraction of the land area, but they are often obscured by clouds. Over the course of the seasons, some changes in the storm patterns and cloud belts on Earth would be observed. Also prominent would be the growth and recession of the winter snowcap across land areas of the Northern Hemisphere.

Scientists have applied the full battery of modern instrumentation to studying Earth in ways not yet been possible for the other planets; thus, much more is known about its structure and composition. It is convenient to consider separate parts of Earth in terms of concentric, roughly spherical layers. Extending from the interior outward, these are the core, the mantle, the crust (including the rocky surface), the hydrosphere (predominantly the oceans, which fill in low places in the crust), the atmosphere (itself divided into spherical zones such as the troposphere, where weather occurs, and the stratosphere, where lies the ozone layer that shields Earth's surface and its organisms against the Sun's ultraviolet rays), and the magnetosphere (an enormous region in space where Earth's magnetic field dominates the behaviour of electrically charged particles from the Sun).

Knowledge about these divisions is summarized in this astronomically oriented overview. The discussion complements other treatments oriented to the Earth sciences and life sciences. The interior, surface, and physical and chemical properties of the planet are discussed in detail in the article EARTH, THE. The geologic and biological development of Earth, including its surface features and the processes by which they are created and modified, are discussed in GEOCHRONOLOGY, CONTINENTAL LANDFORMS, GEOMORPHIC PROCESSES, and PLATE TECTONICS. The behaviour of the atmosphere and of its tenuous, ionized outer reaches is treated in ATMOSPHERE, while the water cycle and major hydrologic features are described in HYDROSPHERE, OCEANS, and RIVERS. The global ecosystem of living organisms and their life-supporting stratum are detailed in BIOSPHERE.

Table 4: Planetary Data for Earth

Mean distance from Sun	149,600,000 km (1.0 AU)
Eccentricity of orbit	0.0167
Inclination of orbit to ecliptic	0.000°
Earth year (sidereal period of revolution)	365.256 days
Rotation period (Earth sidereal day)	23.9345 hours of mean solar time
Earth mean solar day	24.0657 hours of mean sidereal time
Mean orbital velocity	29.79 km/s
Inclination of Equator to orbit	23.45°
Mass	5.976×10^{24} kg
Equatorial radius	6,378.14 km
Polar radius	6,356.78 km
Density	5.52 g/cm³
Mean surface gravity	980 cm/s²
Escape velocity	11.2 km/s
Atmospheric composition	78% molecular nitrogen, 21% molecular oxygen, 0.93% argon, 0.037% carbon dioxide (presently rising), about 1% water (variable)
Mean surface pressure	1 bar
Mean surface temperature	288 K (59° F, 15° C)
Magnetic field strength at Equator	0.3 gauss (but weakening)
Dipole moment	7.9×10^{25} gauss/cm³
Tilt angle of magnetic axis	11.5°
Number of known moons	1 (the Moon)

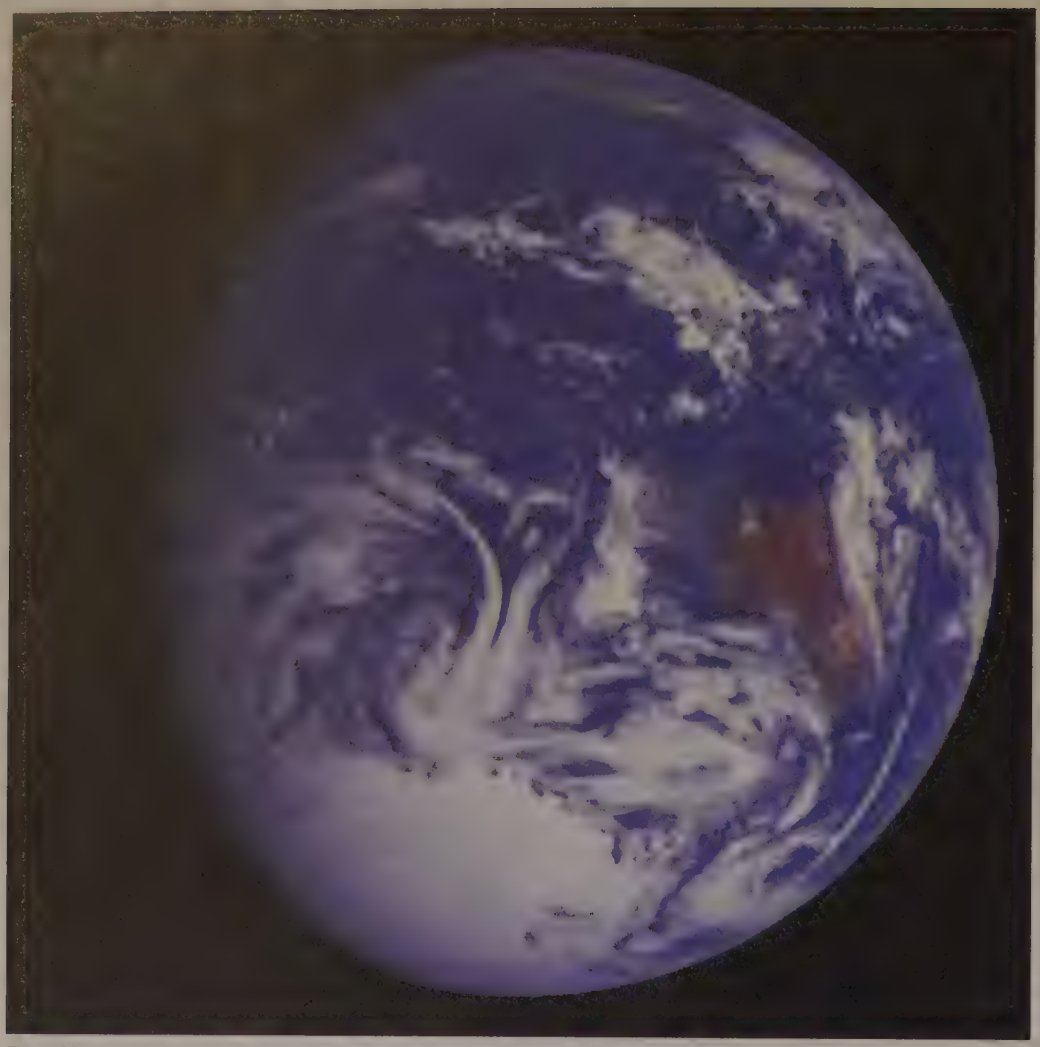


Figure 25: The planet Earth, as imaged by the Galileo spacecraft during its December 1990 flyby en route to Jupiter. The predominance of water on Earth is apparent, both as ocean and in the form of swirling clouds. The landmass at centre right is Australia, and the bright white patch at the bottom is the South Polar ice cap covering Antarctica.
NASA/JPL

BASIC PLANETARY DATA

The mean distance of Earth from the Sun is about 150 million kilometres. The planet orbits the Sun in a path that is presently more nearly a circle (less eccentric) than are the orbits of all but two of the other planets, Venus and Neptune. Earth makes one revolution, or one complete orbit of the Sun, in about 365.25 days. The direction of revolution—counterclockwise as viewed down from the north—is in the same sense, or direction, as the rotation of the Sun; Earth's spin, or rotation about its axis, is also in the same sense, which is called direct or prograde. The rotation period, or length of a sidereal day—23 hours, 56 minutes, and 4 seconds—is similar to that of Mars. Jupiter and most asteroids have days less than half as long, while Mercury and Venus have days more nearly comparable to their orbital periods. The 23.5° tilt, or inclination, of Earth's axis to its orbital plane, also typical, results in greater heating and more hours of daylight in one hemisphere or the other over the course of a year and so is responsible for the cyclic change of seasons.

Axial tilt as reason for seasons

With an equatorial radius of 6,378 kilometres, Earth is the largest of the four inner, terrestrial (rocky) planets, but it is considerably smaller than the gas giants of the outer solar system. Earth has a single satellite, the Moon, which orbits the planet at a mean distance of about 384,400 kilometres. The Moon is one of the bigger natural satellites in the solar system; only the giant planets have moons comparable or larger in size. Some planetary astronomers consider the Earth-Moon system a double planet, with some similarity in that regard to Pluto and its moon, Charon.

Earth's gravitational field is manifested as the attractive force acting on a free body at rest, causing it to accelerate in the general direction of the centre of the planet. Departures from the spherical shape and the effect of Earth's rotation cause gravity to vary with latitude over the terrestrial surface. The average gravitational acceleration at sea level is about 980 centimetres per second per second (32.2 feet per second per second).

Earth's gravity keeps the Moon in its orbit around the planet and also generates tides in the solid body of the Moon. Such deformations are manifested in the form of slight bulges at the lunar surface, detectable only by sensitive instruments. In turn, the Moon's mass exerts a gravitational force that causes tides on Earth. The Sun, much more distant but vastly more massive, also raises tides on Earth. The cyclic movement of the water throughout the ocean basins as a result of the tides (as well as, to a lesser extent, the tidal distortion of the solid Earth) dissipates orbital kinetic energy as heat, producing a gradual slowing of Earth's rotation and a spiraling outward of the Moon's orbit. Currently this slowing lengthens the day by a few thousandths of a second per century. For additional orbital and physical data, see Table 4.

Gradual increase in length of day

THE ATMOSPHERE AND HYDROSPHERE

The blankets of volatile gases and liquids near and above the surface of Earth are of prime importance, along with solar energy, to the sustenance of life on Earth. They are

distributed and recycled throughout the atmosphere and hydrosphere of the planet.

The atmosphere. Earth is surrounded by a relatively thin atmosphere (commonly called air) consisting of a mixture of gases, primarily molecular nitrogen (78 percent) and molecular oxygen (21 percent). Also present are much smaller amounts of gases such as argon (nearly 1 percent), water vapour (averaging 1 percent but highly variable in time and location), carbon dioxide (0.037 percent [370 parts per million] and presently rising), methane (0.00015 percent [1.5 parts per million]), and others, along with minute solid and liquid particles in suspension.

Because Earth has a weak gravitational field (by virtue of its size) and warm atmospheric temperatures (due to its proximity to the Sun) compared with the giant planets, it lacks the most common gases in the universe that they possess: hydrogen and helium. Whereas both the Sun and Jupiter are composed predominantly of these two elements, they could not be retained long on early Earth and rapidly evaporated into interplanetary space. The high oxygen content of Earth's atmosphere is out of the ordinary. Oxygen is a highly reactive gas that, under most planetary conditions, would have combined with other chemicals in the atmosphere, surface, and crust. It is in fact supplied continuously by biological processes; without life, there would be virtually no free oxygen. The 1.5 parts per million of methane in the atmosphere is also far out of chemical equilibrium with the atmosphere and crust: it, too, is of biological origin, with the contribution by human activities far outweighing others.

Anomalous oxygen content

The gases of the atmosphere extend from the surface of Earth to heights of thousands of kilometres, eventually merging with the solar wind—a stream of charged particles that flows outward from the outermost regions of the Sun. The composition of the atmosphere is more or less constant with height to an altitude of about 100 kilometres, with particular exceptions being water vapour and ozone.

The atmosphere is commonly described in terms of distinct layers, or regions. Most of the atmosphere is concentrated in the troposphere, which extends from the surface to an altitude of about 10–15 kilometres, depending on latitude and season. The behaviour of the gases in this layer is controlled by convection. This process involves the turbulent, overturning motions resulting from buoyancy of near-surface air warmed by the Sun. Convection maintains a decreasing vertical temperature gradient—*i.e.*, a temperature decline with altitude—of roughly 6° C (10.8° F) per kilometre through the troposphere. At the top of the troposphere, which is called the tropopause, temperatures have fallen to about −80° C (−112° F). The troposphere is the region where nearly all water vapour exists and essentially all weather occurs.

The dry, tenuous stratosphere lies above the troposphere and extends to an altitude of about 50 kilometres. Convective motions are weak or absent in the stratosphere; motions instead tend to be horizontally oriented. The temperature in this layer increases with altitude. In the upper stratospheric regions, absorption of ultraviolet light from the Sun breaks down molecular oxygen (O_2); recombination of single oxygen atoms with O_2 into ozone (O_3) creates the shielding ozone layer.

Stratospheric ozone layer

Above the relatively warm stratopause is the even more tenuous mesosphere, in which temperatures again decline with altitude to 80–90 kilometres above the surface, where the mesopause is defined. The minimum temperature attained there is extremely variable with season. Temperatures then rise with increasing height through the overlying layer known as the thermosphere. Also above about 80–90 kilometres there is an increasing fraction of charged, or ionized, particles; from this altitude upward defines the ionosphere. Spectacular auroras are generated in this region, particularly along approximately circular zones around the poles, by the interaction of nitrogen and oxygen atoms in the atmosphere with episodic bursts of energetic particles coming from the Sun.

Earth's general atmospheric circulation is driven by the energy of sunlight, which is more abundant in equatorial latitudes. Movement of this heat toward the poles is strongly affected by Earth's rapid rotation and the associated Coriolis force at latitudes away from the Equator (which adds an east-west component to the direction of the winds), resulting in multiple cells of circulating air in each hemisphere. Instabilities (perturbations in the atmospheric flow that grow with time) produce the characteristic high-pressure areas and low-pressure storms of the midlatitudes as well as the fast, eastward-moving jet streams of the upper troposphere that guide the paths of storms. The oceans are massive reservoirs of heat that act largely to smooth out variations in Earth's global temperatures, but their slowly changing currents and temperatures also influence weather and climate, as in the El Niño/Southern Oscillation weather phenomenon (see OCEANS: *Impact of ocean-atmosphere interactions on weather and climate: El Niño/Southern Oscillation and climatic change*).

Earth's atmosphere is not a static feature of the environment. Rather its composition has evolved over geologic time in concert with life and is changing more rapidly today in response to human activities. Roughly halfway through the history of Earth, the atmosphere's unusually high abundance of free oxygen began to develop, through photosynthesis by cyanobacteria (blue-green algae) and saturation of natural surface sinks of oxygen (*e.g.*, relatively oxygen-poor minerals and hydrogen-rich gases exuded from volcanoes). Accumulation of oxygen made it possible for complex cells, which consume oxygen during metabolism and of which all plants and animals are composed, to develop.

Earth's climate at any location varies with the seasons, but there are also longer-term variations in global climate. Volcanic explosions, such as the 1991 eruption of Mount Pinatubo in the Philippines, can inject great quantities of dust particles into the stratosphere, which remain suspended for years, decreasing atmospheric transparency and resulting in measurable cooling worldwide. Much rarer, giant impacts of asteroids and comets can produce even more profound effects, including severe reductions in sunlight for months or years, such as many scientists believe led to the mass extinction of living species at the end of the Cretaceous Period, 65 million years ago. The dominant climate variations observed in the recent geologic record are the ice ages, which are linked to variations in Earth's tilt and its orbital geometry with respect to the Sun.

Longer-term climate variations

The physics of hydrogen fusion leads astronomers to conclude that the Sun was 30 percent less luminous during the earliest history of Earth than it is today. Hence, all else being equal, the oceans should have been frozen. Observations of Earth's planetary neighbours, Mars and Venus, and estimates of the carbon locked in Earth's crust at present suggest that there was much more carbon dioxide in Earth's atmosphere during earlier periods. This would have enhanced warming of the surface via the greenhouse effect and so allowed the oceans to remain liquid. Today there is 100,000 times more carbon dioxide buried in carbonate rocks in Earth's crust than in the atmosphere, in sharp contrast to Venus, whose atmospheric evolution followed a different course.

Between the late 1950s and the end of the 20th century, the amount of carbon dioxide in Earth's atmosphere increased by more than 15 percent because of the burning of fossil fuels (*e.g.*, coal, oil, and natural gas) and the destruction of tropical rainforests, such as that of the Amazon River basin. Computer models predict that a net doubling of carbon dioxide by the middle of the 21st century could lead to a global warming of 1.5–4.5° C (2.7–8.1° F) averaged over the planet, which would have profound effects on sea level and agriculture. Although this conclusion has been criticized by some on the basis that the warming observed so far has not kept pace with the projection, analyses of ocean temperature data have suggested that much of the warming during the 20th century actually occurred in the oceans themselves—and will eventually appear in the atmosphere.

Rising atmospheric carbon dioxide

Another present concern regarding the atmosphere is the impact of human activities on the stratospheric ozone layer. Chemical reactions involving traces of man-made chlorofluorocarbons (CFCs) were found in the mid-1980s to be creating temporary holes in the ozone layer, particularly over Antarctica, during polar spring. The discovery of

a growing depletion of ozone over the highly populated temperate latitudes was even more disturbing, because the short-wavelength ultraviolet radiation that the ozone layer effectively absorbs has been found to cause skin cancer. International agreements in place to halt the production of the most egregious ozone-destroying CFCs will eventually halt and reverse the depletion, but only by the middle of the 21st century, because of the long residence time of these chemicals in the stratosphere.

The hydrosphere. Earth's hydrosphere is a discontinuous layer of water at or near the planet's surface; it includes all liquid and frozen surface waters, groundwater held in soil and rock, and atmospheric water vapour. Unique within the solar system, the hydrosphere is essential to all life as it is presently understood. Almost 71 percent of Earth's surface is covered by saltwater oceans, with a volume of about 1.4 billion cubic kilometres (336 million cubic miles) and an average temperature of about 4° C (39.2° F), not far above the freezing point of water. The oceans contain about 97 percent of the planet's water volume. The remainder occurs as fresh water, three-quarters of which is locked up in the form of ice at polar latitudes. Most of the remaining fresh water is groundwater held in soils and rocks; less than 1 percent occurs in lakes and rivers. In terms of percentage, atmospheric water vapour is negligible, but the transport of water evaporated from the oceans onto land surfaces is an integral part of the hydrologic cycle that renews and sustains life.

Importance of atmospheric water transport

The hydrologic cycle involves the transfer of water from the oceans through the atmosphere to the continents and back to the oceans over and beneath the land surface. The cycle includes processes such as precipitation, evaporation, transpiration, infiltration, percolation, and runoff. These processes operate throughout the entire hydrosphere, which extends from about 15 kilometres into the atmosphere to roughly 5 kilometres into the crust.

About one-third of the solar energy that reaches Earth's surface is expended on evaporating ocean water. The resulting atmospheric moisture and humidity condense into clouds, rain, snow, and dew. Moisture is a crucial factor in determining weather. It is the driving force behind storms and is responsible for separating electrical charge, which is the cause of lightning and thus of natural wildland fires, which have an important role in some ecosystems. Moisture wets the land, replenishes subterranean aquifers, chemically weathers the rocks, erodes the landscape, nourishes life, and fills the rivers, which carry dissolved chemicals and sediments back into the oceans.

Water also plays a vital role in the carbon dioxide cycle (a part of the more inclusive carbon cycle). Under the action of water and dissolved carbon dioxide, calcium is weathered from continental rocks and carried to the oceans, where it combines to form calcium carbonates (including shells of marine life). Eventually the carbonates are deposited on the seafloor and are lithified to form limestones. Some of these carbonate rocks are later dragged deep into Earth's interior by the global process of plate tectonics (see below *The outer shell*) and melted, resulting in a rerelease of carbon dioxide (from volcanoes, for example) into the atmosphere. Cyclic processing of water, carbon dioxide, and oxygen through geologic and biological systems on Earth has been fundamental to maintaining the habitability of the planet with time and to shaping the erosion and weathering of the continents, and it contrasts sharply with the lack of such processes on Venus. (On Mars there is evidence of past episodes of liquid water erosion—and possibly limited amounts of such erosion today.)

THE OUTER SHELL

Earth's outermost rigid, rocky layer is called the crust. It is composed of low-density, easily melted rocks; the continental crust is predominantly granitic, while composition of the oceanic crust corresponds mainly to that of basalt and gabbro. Analyses of seismic waves, generated by earthquakes within Earth's interior, show that the crust extends about 50 kilometres beneath the continents but only 5–10 kilometres beneath the ocean floors.

At the base of the crust, a sharp change in the observed behaviour of seismic waves marks the interface with the mantle. The mantle is composed of denser rocks, on which the rocks of the crust float. On geologic timescales, the mantle behaves as a very viscous fluid and responds to stress by flowing. Together the crust and the uppermost mantle act mechanically as a single rigid layer, called the lithosphere.

The lithospheric outer shell of Earth is not one continuous piece but is broken, like a slightly cracked eggshell, into about a dozen major separate rigid blocks, or plates. There are two types of plates, oceanic and continental. An example of an oceanic plate is the Pacific Plate, which extends from the East Pacific Rise to the deep-sea trenches bordering the western part of the Pacific basin. A continental plate is exemplified by the North American Plate, which includes North America as well as the oceanic crust between it and a portion of the Mid-Atlantic Ridge. The latter is an enormous submarine mountain chain that extends down the axis of the Atlantic basin, passing midway between Africa and North and South America.

The lithospheric plates are about 60 kilometres thick beneath the oceans and 100–200 kilometres beneath the continents. (It should be noted that these thicknesses are defined by the mechanical rigidity of the lithospheric material. They do not correspond to the thickness of the crust, which is defined at its base by the discontinuity in seismic wave behaviour, as cited above.) They ride on a weak, perhaps partially molten, layer of the upper mantle called the asthenosphere. Slow convection currents deep within the mantle generated by radioactive heating of the interior drive lateral movements of the plates (and the continents on top of them) at a rate of several centimetres per year. The plates interact along their margins, and these boundaries are classified into three general types on the basis of the relative motions of the adjacent plates: divergent, convergent, and transform.

Interaction of lithospheric plates

In areas of divergence, two plates move away from each other. Buoyant upwelling motions in the mantle force the plates apart at rift zones (such as along the middle of the Atlantic Ocean floor) where magmas from the underlying mantle rise to form new oceanic crustal rocks.

Lithospheric plates move toward each other along convergent boundaries. When a continental plate and an oceanic plate come together, the leading edge of the oceanic plate is forced beneath the continental plate and into the asthenosphere—a process called subduction. Only the thinner, denser slabs of oceanic crust will subduct. When two thicker, more buoyant continents come together at convergent zones, they tend to buckle, producing great mountain ranges. The Himalayas, along with the adjacent Plateau of Tibet, were formed during such a collision, when India was carried into the Eurasian Plate by relative motion of the Indian-Australian Plate.

At the third type of plate boundary, the transform (or strike-slip) variety, two plates slide parallel to one another in opposite directions. These areas are often associated with high seismicity, as stresses that build up in the sliding crustal slabs are released at intervals to generate earthquakes. The San Andreas Fault in California is an example of this type of boundary, which is also known as a fault or fracture zone.

Most of Earth's active tectonic processes, including nearly all earthquakes, occur near plate margins. Volcanoes form along zones of subduction, because the oceanic crust tends to be remelted as it descends into the hot mantle and then rises to the surface as lava. Chains of active, often explosive volcanoes are thus formed in such places as the western Pacific and the west coasts of the Americas. Older mountain ranges, eroded by weathering and runoff, mark zones of earlier plate-margin activity.

Oldest regions of Earth

The oldest, most geologically stable parts of Earth are the central cores of some continents (such as Australia, parts of Africa, and northern North America). Called continental shields, they are regions where mountain building, faulting, and other tectonic processes are diminished compared with the activity that occurs at the boundaries between plates. Because of their stability, erosion has had the time to flatten the topography of continental shields. It is also on the shields that geologic evidence of crater scars from ancient impacts of asteroids and comets is better pre-

served. Even there, however, tectonic processes and the action of water have erased many ancient features. In contrast, much of the oceanic crust is substantially younger (tens of millions of years old), and none dates back more than 200 million years.

Plate tectonics

This conceptual framework in which scientists now understand the evolution of Earth's lithosphere—termed plate tectonics—is almost universally accepted, although many details remain to be worked out. For example, scientists have yet to reach a general agreement as to when the original continental cores formed or how long ago modern plate-tectonic processes began to operate.

Once major continental shields grew, plate tectonics was characterized by the cyclic assembly and breakup of supercontinents created by the amalgamation of many smaller continental cores and island arcs. Scientists have identified two such cycles in the geologic record. A supercontinent began breaking up about 700 million years ago, in late Precambrian time, into several major continents, but by about 250 million years ago, near the beginning of the Triassic Period, the continued drifting of these continents resulted in their fusion again into a single supercontinent called Pangaea. Some 70 million years later, Pangaea began to fragment, giving rise to today's continental configuration. The distribution is still asymmetric, with continents predominantly located in the Northern Hemisphere opposite the Pacific basin.

Of the four terrestrial planets, only Earth shows evidence of long-term, pervasive plate tectonics. Both Venus and Mars exhibit geology dominated by basaltic volcanism on a largely immovable crust, with only faint hints of possibly limited episodes of horizontal plate motion. Mercury is intrinsically much denser than the other terrestrial planets, which implies a larger metallic core; its surface is mostly covered with impact craters, but it also shows a global pattern of scarps suggesting shrinkage of the planet, associated perhaps with interior cooling. Apparently essential to the kind of plate tectonics that occurs on Earth are large planetary size (hence, high heat flow and thin crust), which eliminates Mercury and Mars, and pervasive crustal water to soften the rock, which Venus lost very early in its history.

THE INTERIOR

More than 90 percent of Earth's mass is composed of iron, oxygen, silicon, and magnesium, elements that can form the crystalline minerals known as silicates. Nevertheless, in chemical and mineralogical composition, as in physical properties, Earth is far from homogeneous. Apart from the superficial lateral differences near the surface (*i.e.*, in the compositions of the continental and oceanic crusts), Earth's principal differences vary with distance toward the centre. This is due to increasing temperatures and pressures and to the original segregation of materials, soon after Earth accreted from the solar nebula about 4.56 billion years ago, into a metal-rich core, a silicate-rich mantle, and the more highly refined crustal rocks. Earth is geochemically differentiated to a great extent. Crustal rocks contain several times as much of the rock-forming element aluminum as does the rest of the solid Earth and many dozens of times as much uranium. On the other hand, the crust, which accounts for a mere 0.4 percent of Earth's mass, contains less than 0.1 percent of its iron. Between 85 and 90 percent of Earth's iron is concentrated in the core.

The increasing pressure with depth causes phase changes in crustal rocks at depths between 5 and 50 kilometres, which marks the top of the upper mantle, as mentioned above. This transition area is called the Mohorovičić discontinuity, or Moho. Most basaltic magmas are generated in the upper mantle at depths of hundreds of kilometres. The upper mantle, which is rich in the olivine, pyroxene, and silicate perovskite minerals, shows significant lateral differences in composition. A large fraction of Earth's interior, from a depth of about 650 kilometres down to 2,900 kilometres, consists of the lower mantle, which is composed chiefly of magnesium- and iron-bearing silicates, including the high-pressure equivalents of olivine and pyroxene.

Transition from crust to mantle

The mantle is not static but rather churns slowly in convective motions, with hotter material rising up and cooler material sinking; through this process, Earth gradually loses its internal heat. In addition to being the driving force of horizontal plate motion, mantle convection is manifested in the occurrence of temporary superplumes—huge, rising jets of hot, partially molten rock—which may originate from a deep layer near the core-mantle interface. Much larger than ordinary thermal plumes, such as that associated with the Hawaiian island chain in the central Pacific, superplumes may have had profound effects on Earth's geologic history and even on its climate. One outburst of global volcanism about 65 million years ago, which created the vast flood basalt deposits known as the Deccan Traps on the Indian subcontinent, may have been associated with a superplume, though this model is far from universally accepted.

Effects of mantle superplumes

With a radius of almost 3,500 kilometres, Earth's core is about the size of the entire planet Mars. About one-third of Earth's mass is contained in the core, most of which is liquid iron alloyed with some lighter, cosmically abundant components (*e.g.*, sulfur, oxygen, and even, controversially, hydrogen). Its liquid nature is revealed by the failure of shear-type seismic waves to penetrate the core. A small, central part of the core, however, below a depth of about 5,100 kilometres, is solid. Temperatures in the core are extremely hot, ranging from 4,000–5,000 K (roughly 6,700–8,500° F, 3,700–4,700° C) at the outer part of the core to 5,000–7,000 K (8,500–12,100° F, 4,700–6,700° C) in the centre, comparable to the surface of the Sun. The core's reservoir of heat may contribute as much as one-fifth of all the internal heat that ultimately flows to the surface of Earth.

THE GEOMAGNETIC FIELD AND MAGNETOSPHERE

Helical fluid motions in Earth's electrically conducting liquid outer core have a magnetohydrodynamic dynamo effect, giving rise to the geomagnetic field. The planet's sizable, hot core, along with its rapid spin, probably accounts for the exceptional strength of the magnetic field of Earth compared with those of the other terrestrial planets. Venus, for example, which has a metallic core that may be similar to Earth's in size, rotates very slowly and has no detected intrinsic magnetic field. Mercury and Mars have only small intrinsic magnetic fields.

Earth's main magnetic field permeates the planet and an enormous volume of space surrounding it. A great teardrop-shaped region of space called the magnetosphere is formed by the interaction of Earth's field with the solar wind. At a distance of about 65,000 outward toward the Sun, the pressure of the solar wind is balanced by the geomagnetic field. This serves as an obstacle to the solar wind, and the flow of charged particles, or plasma, is deflected around Earth by the resulting bow shock. The magnetosphere so produced streams out into an elongated magnetotail that stretches several million kilometres downstream from Earth away from the Sun.

Plasma particles from the solar wind can leak through the magnetopause, the sunward boundary of the magnetosphere, and populate its interior; charged particles from the Earth's ionosphere also enter the magnetosphere. The magnetotail can store for hours an enormous amount of energy—several billion megajoules, which is roughly equivalent to the yearly electricity production of many smaller countries). The energy is released in dynamic structural reconfigurations of the magnetosphere, called geomagnetic substorms, which often result in the precipitation of energetic particles into the ionosphere, giving rise to fluorescing auroral displays.

Converging magnetic field lines fairly close to Earth trap highly energetic particles so that they gyrate between the Northern and Southern hemispheres and slowly drift longitudinally around the planet in two concentric doughnut-shaped zones, known as the Van Allen radiation belts. Many of the charged particles trapped in these belts are produced when energetic cosmic rays strike Earth's upper atmosphere, producing neutrons that then decay into electrons, which are negatively charged, and protons, which are positively charged. Others come from the solar wind or Earth's atmosphere. Earth's magnetosphere has been ex-

Van Allen belts

tensively studied ever since the discovery of the Van Allen belts in the late 1950s, and space physicists have extended their studies of plasma processes to the vicinities of comets and other planets. (C.R.C/J.I.L.)

The Moon

The Moon, designated ☾ in astronomy, is Earth's sole natural satellite and nearest large celestial body. Known since prehistoric times, it is the brightest object in the sky after the Sun. Its name in English, like that of Earth, is of Germanic and Old English derivation.

Centuries of observation and scientific investigation have been centred on the nature and origin of the Moon. Early studies of the Moon's motion and position allowed the prediction of tides and led to the development of calendars. The Moon was the first new world on which humans set foot; the information brought back from those expeditions, together with that collected by automated spacecraft and remote-sensing observations, have led to a knowledge of the Moon that surpasses that of any other cosmic body except Earth itself. Although many questions remain about its composition, structure, and history, it has become clear that the Moon holds keys to understanding the origin of Earth and the solar system. Moreover, given its nearness to Earth, its rich potential as a source of materials and energy, and its qualifications as a laboratory for planetary science and a place to learn how to live and work in space for extended times, the Moon remains a prime location for humankind's first settlements beyond Earth orbit.

DISTINCTIVE FEATURES

The Moon is a spherical rocky body, probably with a small metallic core, revolving around Earth in a slightly eccentric orbit at a mean distance of about 384,000 kilometres. Its equatorial radius is 1,738 kilometres, and its shape is slightly flattened in a such a way that it bulges a little in the direction of Earth. Its mass distribution is not uniform—the centre of mass is displaced about 2 kilometres toward Earth relative to the centre of the lunar sphere, and it also has surface mass concentrations, called mascons, that cause the Moon's gravitational field to increase over local areas. The Moon has no global magnetic field like that of Earth, but some of its surface rocks have remanent magnetism, indicating one or more periods of magnetic activity in the past. The Moon presently has very slight seismic activity and little heat flow from the interior, indications that most internal activity ceased long ago.

Scientists now believe that more than four billion years ago the Moon was subject to violent heating—probably from its formation—which resulted in its differentiation, or chemical separation, into a less dense crust and a more dense underlying mantle. This was followed hundreds of millions of years later by a second episode of heating—this time from internal radioactivity—which resulted in volcanic outpourings of lava.

The Moon's mean density is 3.34 grams per cubic centimetre, close to that of Earth's mantle. Because of the Moon's small size and mass, its surface gravity is only about one-sixth of the planet's; it retains so little atmosphere that the molecules of any gases present on the surface move without collision. In the absence of an atmospheric shield to protect the surface from bombardment, countless bodies ranging in size from asteroids to tiny particles have struck and cratered the Moon. This has formed a debris layer, or regolith, consisting of rock fragments of all sizes down to the finest dust. In the ancient past the largest impacts made great basins, some of which were later partly filled by the enormous lava floods. These great dark plains, called maria (singular *mare* [Latin: "sea"]), are clearly visible to the naked eye from Earth. The dark maria and the lighter highlands, whose unchanging patterns many people recognize as the "man in the moon," constitute the two main kinds of lunar territory. The mascons are regions where particularly dense lavas rose up from the mantle and flooded into basins. Lunar mountains, located mostly along the rims of ancient basins, are tall but not steep or sharp-peaked, because all lunar landforms have been eroded by the unending rain

Maria and highlands

NASA/JPL/Caltech

Figure 26: The familiar near side of Earth's Moon, photographed on Dec. 7, 1992, by the Galileo spacecraft on its way to Jupiter. Two primary kinds of terrain are visible—the lighter areas, which constitute the heavily cratered and very old highlands, and the darker, roughly circular plains, traditionally called maria, which are relatively young lava-filled impact basins.

of impacts. For additional orbital and physical data, see Table 5.

Table 5: Properties of the Moon and the Earth-Moon System

	Moon	Earth
Mean distance from Earth	384,400 km	—
Mass	0.0735×10^{24} kg	5.976×10^{24} kg
Equatorial radius	1,738 km	6,378 km
Surface area	37.9×10^{6} km^2	509.6×10^{6} km^2 (land = 148×10^{6} km^2)
Mean density	3.34 g/cm^3	5.52 g/cm^3
Mean surface gravity	162 cm/s^2	980 cm/s^2
Escape velocity	2.38 km/s	11.2 km/s
Period of orbit around Earth (sidereal period of revolution)	27.3217 Earth days	—
Rotation period	synchronous with orbital period	23.9345 hours
Mean surface temperature	day 380 K (224° F, 107° C), night 120 K (−224° F, 153° C)	288 K (59° F, 15° C)
Temperature extremes	396 K (253° F, 123° C) to 40 K (−388° F, −233° C)	331 K (136° F, 58° C) to 184 K (−128° F, −89° C)
Average heat flow	29 mW/m^2	63 mW/m^2

PRINCIPAL CHARACTERISTICS OF THE EARTH-MOON SYSTEM

In addition to its nearness to Earth, the Moon is relatively massive compared with the planet—the ratio of their masses is much larger than those of other natural satellites to the planets that they orbit, with the exception of Charon and Pluto. The Moon and Earth consequently exert a strong gravitational influence on each other, forming a system having distinct properties and behaviour of its own. Table 5 compares some salient characteristics of the two bodies.

Although the Moon is commonly described as orbiting Earth, it is more accurate to say that the two bodies orbit each other about a common centre of mass. Called the barycentre, this point lies inside Earth about 4,700 kilometres from its centre. Also more accurately, it is the barycentre, rather than the centre of Earth, that follows an elliptical path around the Sun in accord with Kepler's laws of planetary motion. The orbital geometry of the Moon, Earth, and the Sun gives rise to the Moon's phases and to the phenomena of lunar and solar eclipses. The configuration and motions of the Earth-Moon system are illustrated in Figure 28.

The distance between the Moon and Earth varies rather widely because of the combined gravity of Earth, the Sun, and the planets. For example, in the last three decades of the 20th century, the Moon's apogee—the farthest distance that it travels from Earth in a revolution—ranged between about 404,000 and 406,700 kilometres, while its perigee—the closest that it comes to Earth—ranged between about 356,500 and 370,400 kilometres. Tidal interactions, the cyclic deformations in each body caused by the gravitational attraction of the other, have braked the Moon's spin such that it now rotates at the same rate as it revolves around Earth, thus always keeping the same side facing the planet. As discovered by the Italian-born French astronomer Gian Domenico Cassini in 1692, the Moon's spin axis precesses with respect to its orbital plane; *i.e.*, its orientation changes slowly over time, tracing out a circular path.

In accord with Kepler's second law, the eccentricity of the Moon's orbit results in its traveling faster in that part of its orbit nearer Earth and slower in the part farther away. Combined with the Moon's constant spin rate, these changes in speed give rise to an apparent oscillation, or libration, which over time allows an observer on Earth to see more than half of the lunar surface. In addition to this apparent turning motion, the Moon actually does rock slightly to and fro in both longitude and latitude, and the observer's vantage point moves with Earth's rotation. As a result of all these motions, more than 59 percent of the lunar surface can be seen at one time or another from Earth.

Lunar libration

The orbital eccentricity also affects solar eclipses, in which the Moon passes between the Sun and Earth, casting a moving shadow across Earth's sunlit surface. If a solar eclipse occurs when the Moon is near perigee, ob-

F.J. Doyle/National Space Science and Data Center

Figure 27: View of the Moon never seen from Earth, predominantly the heavily cratered far side, photographed by Apollo 16 astronauts in April 1972. The near-side impact basin Mare Crisium is the large, dark marking on the upper left limb. Although the far side is well scarred with giant basins, these never filled with lava to form maria.

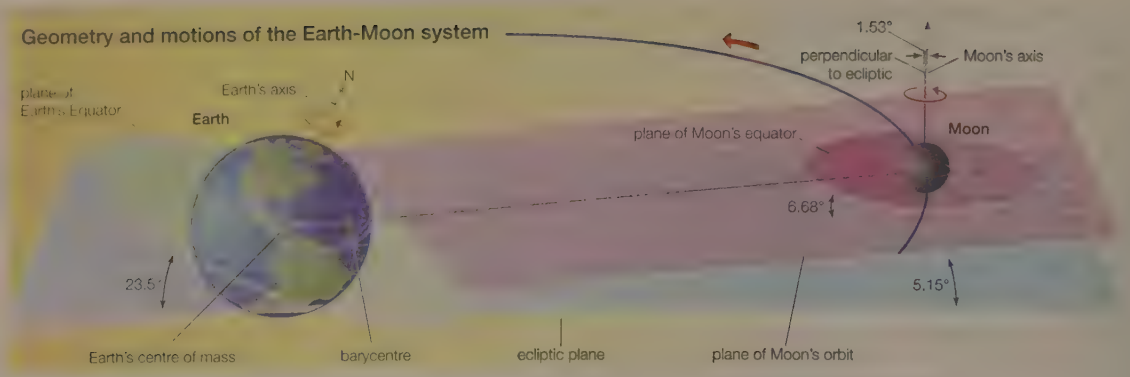


Figure 28: Relative configuration and motions of Earth and the Moon.

servers along the path of the Moon's dark inner shadow (umbra) see a total eclipse. If the Moon is near apogee, it does not quite cover the Sun; the resulting eclipse is annular, and observers can see a thin ring of the solar disk around the Moon's silhouette.

The Moon and Earth presently orbit the barycentre in 27.322 days, the sidereal month, or sidereal revolution period of the Moon. Because the whole system is moving around the Sun once per year, the angle of illumination changes about one degree per day, so that the time from one full moon to the next is 29.531 days, the synodic month, or synodic revolution period of the Moon. As a result, the Moon's terminator—the dividing line between dayside and nightside—moves once around the Moon in this synodic period, exposing most locations to alternating periods of sunlight and darkness each nearly 15 Earth days long. The sidereal and synodic periods are slowly changing with time because of tidal interactions. Although tidal friction is slowing Earth's rotation, conservation of momentum dictates that the angular momentum of the Earth-Moon system stays constant. Consequently, the Moon is slowly receding from Earth, with the result that both the day and the month are getting longer. Extending this relationship into the past, both periods must have been significantly shorter hundreds of millions of years ago—a hypothesis confirmed from measurements of the daily and tide-related growth rings of fossil corals.

Slow recession from Earth

Because the Moon's spin axis is almost perpendicular to the plane of the ecliptic (the plane of Earth's orbit around the Sun)—inclined only 1½ degrees from the vertical—the Moon has no seasons. Sunlight is always nearly horizontal at the lunar poles, resulting in permanently cold and dark environments at the bottoms of deep craters.

MOTIONS OF THE MOON

The study of the Moon's motions has been central to the growth of knowledge not only about the Moon itself but also about fundamentals of celestial mechanics and physics. As the stars appear to move westward because of Earth's daily rotation and its annual motion about the Sun, so the Moon slowly moves eastward, rising later each day and passing through its phases: new, first quarter, full, last quarter, and new again each month. From ancient times, investigators have looked for small departures from the motions predicted. The English physicist Isaac Newton used lunar observations in developing his theory of gravitation in the late 17th century, and he was able to show some effects of solar gravity in perturbing the Moon's motion. By the 18th and 19th centuries the mathematical study of lunar movements, both orbital and rotational, was advancing, driven in part by the need for precise tables of the predicted positions of celestial bodies (ephemerides) for navigation. While theory developed with improved observations, small and puzzling discrepancies continued to appear. It gradually became evident that some arise from irregularities in Earth's rotation rate, others from minor tidal effects on Earth and the Moon.

Space exploration brought a need for greatly increased accuracy, and at the same time the availability of fast computers and new observational tools provided the means for attaining it. Optical and radio observations vastly improved—for example, retroreflectors placed on the lunar surface by Apollo astronauts allowed laser ranging of the Moon from Earth, and new techniques of radio astronomy permitted observations of celestial radio sources as the Moon occulted them. These observations, having precisions on the order of centimetres, have enabled scientists to measure changes in the Moon's speed caused by terrestrial tidal momentum exchange, have advanced understanding of the theories of relativity, and are leading to improved geophysical knowledge of both the Moon and Earth.

THE ATMOSPHERE

Though the Moon is surrounded by a vacuum higher than is usually created in laboratories on Earth, its atmosphere is extensive and of high scientific interest. During the two-week daytime period, atoms and molecules are ejected by a variety of processes from the lunar surface, ionized by the solar wind, and then driven by electromagnetic effects as a collisionless plasma. The position of the Moon in its orbit determines the behaviour of the atmosphere. For part of each month, when the Moon is on the sunward side of Earth, atmospheric gases collide with the undisturbed solar wind; in other parts of the orbit, they move into and out of the elongated tail of Earth's magnetosphere, an enormous region of space where the planet's magnetic field dominates the behaviour of electrically charged particles. In addition, the low temperatures on the Moon's nightside and in permanently shaded polar craters provide cold traps for condensable gases.

Effect of Moon's orbital position

Instruments placed on the lunar surface by Apollo astronauts measured various properties of the Moon's atmosphere, but analysis of the data was difficult because the atmosphere's extreme thinness made contamination from Apollo-originated gases a significant factor. The main gases naturally present are neon, hydrogen, helium, and argon. The argon is mostly radiogenic—*i.e.*, it is released from lunar rocks by the decay of radioactive potassium. Lunar night temperatures are low enough for the argon to condense but not the neon, hydrogen, or helium, which originate in the solar wind and remain in the atmosphere as gases unless implanted in soil particles.

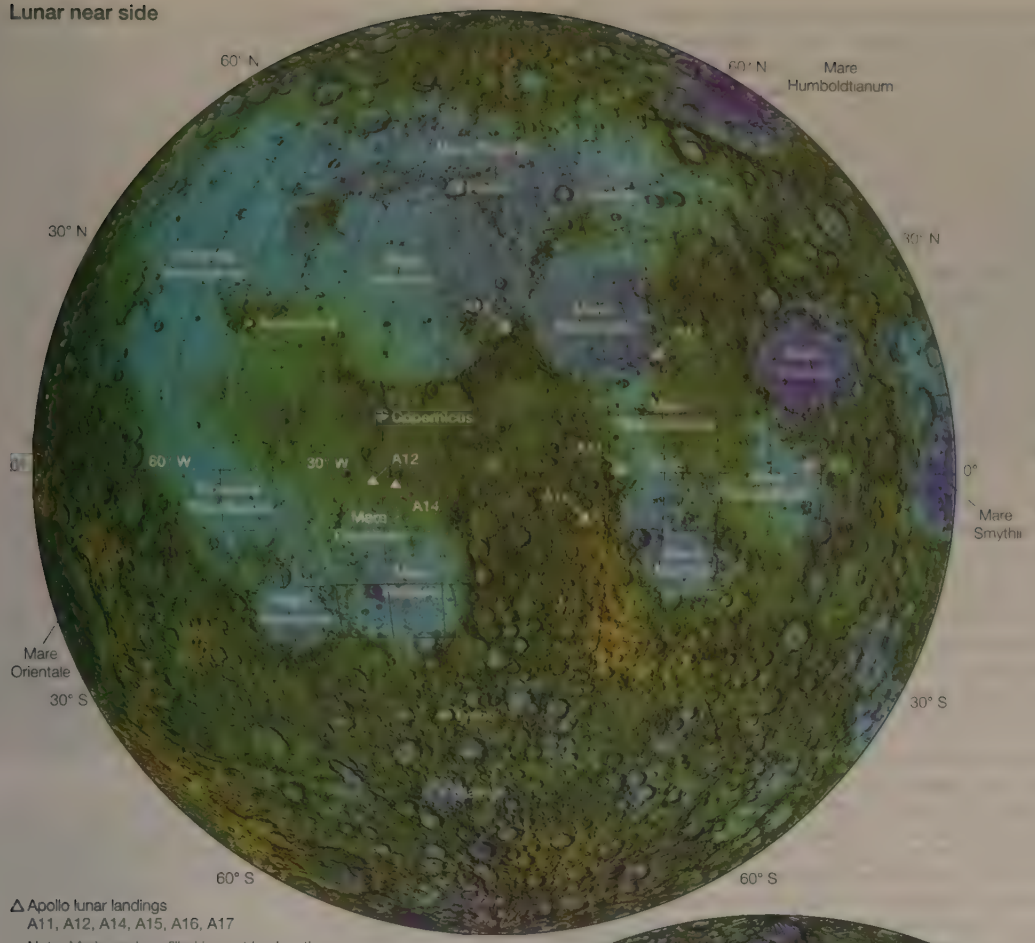
In addition to the near-surface gases and the extensive sodium-potassium cloud detected around the Moon (see below *Effects of impacts and volcanism*), a small amount of dust circulates within a few metres of the lunar surface. This is believed to be suspended electrostatically.

THE LUNAR SURFACE

Large-scale features. With binoculars or a small telescope, an observer can see details of the Moon's near side in addition to the pattern of maria and highlands. As the Moon passes through its phases, the terminator moves slowly across the Moon's disk, its long shadows revealing

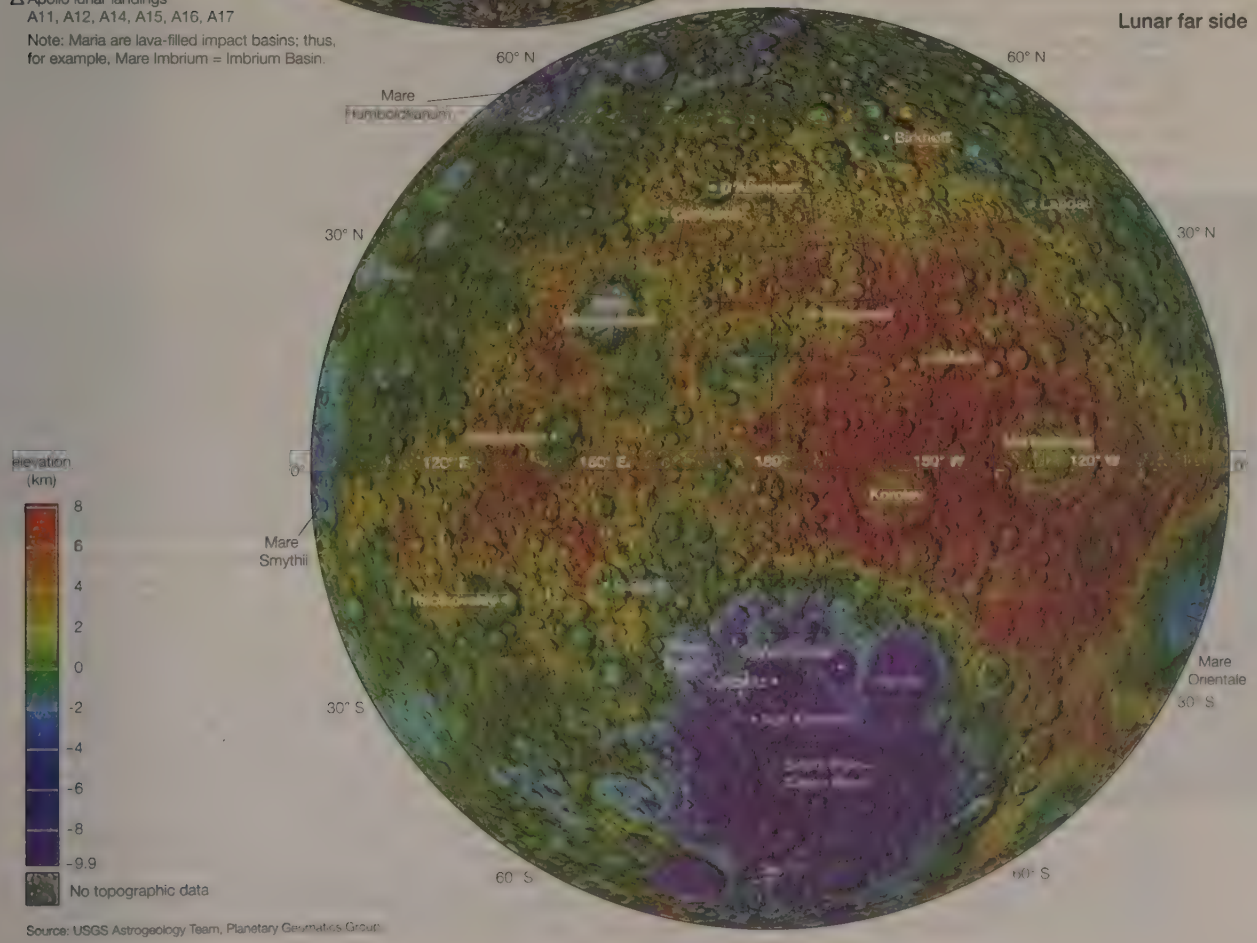
Near-side appearance

Figure 29 (facing page): Topographic maps of the Moon's near- and far-side hemispheres, derived from laser altimetry and imagery data gathered by the Clementine spacecraft from lunar orbit in 1994. Relief is colour-coded according to the key at lower left, with values expressed as distance above or below the Moon's mean radius. Selected major topographic features and Apollo landing sites are labeled.



△ Apollo lunar landings
A11, A12, A14, A15, A16, A17

Note: Maria are lava-filled impact basins; thus, for example, Mare Imbrium = Imbrium Basin.



Source: USGS Astrogeology Team, Planetary Geomatics Group.

the relief of mountains and craters. At full moon the relief disappears, replaced by the contrast between lighter and darker surfaces. Though the full moon is brilliant at night, the Moon is actually a dark object, reflecting only a few percent (albedo 0.07) of the sunlight that strikes it. Beginning with the Italian scientist Galileo's sketches in the early 17th century and continuing into the 19th century, astronomers mapped and named the visible features down to a resolution of a few kilometres, the best that can be accomplished when viewing the Moon through Earth's turbulent atmosphere. The work culminated in a great, hand-drawn lunar atlas made by observers in Berlin and Athens. This was followed by a lengthy hiatus until the mid-20th century, when it became apparent that human travel to the Moon might eventually be possible. In the 1950s another great atlas was compiled, this time a photographic one published in 1960 under the sponsorship of the U.S. Air Force.

Dominance of impacts

Astronomers long debated over whether the Moon's topographic features had been caused by volcanism. Only in the 20th century did the dominance of impacts in the shaping of the lunar surface become clear. Every highland region is heavily cratered—evidence for repeated collisions with large bodies. (The survival of similar large impact structures on Earth is relatively rare because of Earth's geologic activity and weathering.) The maria, on the other hand, show much less cratering and thus must be significantly younger. Mountains are mostly parts of the upthrust rims of ancient impact basins. Volcanic activity has occurred within the Moon, but the results are mostly quite different from those on Earth. The lavas that upwelled in floods to form the maria were very fluid. Evidence of volcanic mountain building as has occurred on Earth is limited to a few fields of small, low domes.

First views of far side

For millennia people wondered about the appearance of the Moon's unseen side. The mystery began to be dispelled with the flight of the Soviet space probe Luna 3 in 1959, which returned the first photographs of the far side. In contrast to the near side, the surface displayed in the Luna 3 images consisted mostly of highlands, with only small areas of dark mare material. Later missions showed that the ancient far-side highlands are scarred by huge basins but that these basins are not filled with lava.

Effects of impacts and volcanism. The dominant consequences of impacts are observed in every lunar scene. At the largest scale are the ancient basins, which extend hundreds of kilometres across. A beautiful example is Orientale Basin, or Mare Orientale (Figure 30), whose mountain walls can just be seen from Earth near the Moon's limb (the apparent edge of the lunar disk) when the lunar libration is favourable. Its multiring ramparts are characteristic of the largest basins; they are accented by the partial lava flooding of low regions between the rings. Orientale Basin appears to be the youngest large impact basin on the Moon.

NASA/Lunar Planetary Institute

Figure 30: The multiringed Orientale Basin, or Mare Orientale, on the Moon, in a photograph taken in 1967 by Lunar Orbiter 4. The giant impact structure's outermost rim, the Cordillera Mountains, is 930 kilometres in diameter. Orientale is located on the western limb of the lunar near side. Unlike other near-side basins, it is only partially flooded by mare lavas.

Figure 31: Sinuous rilles near the ancient, mostly buried crater Prinz on the Moon, in an image taken by Apollo 15 astronauts in 1971. Such features are believed to be similar to channels cut by lava flows on Earth. The most conspicuous rille in the image, Rima Prinz, appears to originate from a small volcanic crater (upper centre) on the rim of the crater Prinz; it trends westward (right) under the crater rim before turning northward (down).

NASA/Goddard Space Flight Center

Smaller impact features, ranging in diameter from tens of kilometres to microscopic size, are described by the term *crater*. The relative ages of lunar craters are indicated by their form and structural features. Young craters have rugged profiles and are surrounded by hummocky blankets of debris, called ejecta, and long, light-coloured rays made by expelled material hitting the lunar surface. Older craters have rounded and subdued profiles, the result of continued bombardment.

Cratering process

A crater's form and structure also yields information about the impact process. When a body strikes a much larger one at speeds of many kilometres per second, the available kinetic energy is enough to completely melt, even partly vaporize, the impacting body along with a small portion of its target material. On impact, a melt sheet is thrown out, along with quantities of rubble, to form the ejecta blanket around the contact site. Meanwhile a shock travels into the subsurface, shattering mineral structures and leaving a telltale signature in the rocks. The initial cup-shaped cavity is unstable and, depending on its size, evolves in different ways. A typical end result is the great crater Aristarchus, with slumping terraces in its walls and a central peak. Aristarchus is about 40 kilometres wide and 4 kilometres deep.

The region around Aristarchus shows a number of peculiar lunar features, some of which have origins not yet well explained. The Aristarchus impact occurred on an elevated, old-looking surface surrounded by lavas of the northern part of the mare known as Oceanus Procellarum.

Rilles

These lava flows inundated the older crater Prinz, whose rim is now only partly visible. At one point on the rim, an apparently volcanic event produced a crater; subsequently a long, winding channel, called a sinuous rille, emerged to flow across the mare (Figure 31). Other sinuous rilles are found nearby, including the largest one on the Moon, discovered by the German astronomer Johann Schröter in 1787. Named in his honour, Schröter's Valley is a deep, winding channel, hundreds of kilometres long, with a smaller inner channel that meanders just as slow rivers do on Earth. The end of this "river" simply tapers away to nothing and disappears on the mare plains. In some way that remains to be accounted for, hundreds of cubic kilometres of fluid and excavated mare material vanished.

The results of seismic and heat-flow measurements suggest that any volcanic activity that persists on the Moon is slight by comparison with that of Earth. Over the years, reliable observers have reported seeing transient events of a possibly volcanic nature, and some spectroscopic evidence for them exists. The question of whether the Moon is volcanically active remains open.

Telescopic observers beginning in the 19th century applied the term *rille* to several types of trenchlike lunar features. In addition to sinuous rilles, there are straight and branching rilles that appear to be tension cracks, and some of these—such as Rima Hyginus and the rilles around the floor of the large old crater Alphonsus—are peppered with rimless eruption craters. Though the Moon shows both tension and compression features (low wrinkle ridges, usually near mare margins, may result from compression), it gives no evidence of having experienced the large, lateral motions of plate tectonics marked by faults in Earth's crust.

L.J. Kosofsky/National Space Science Data Center

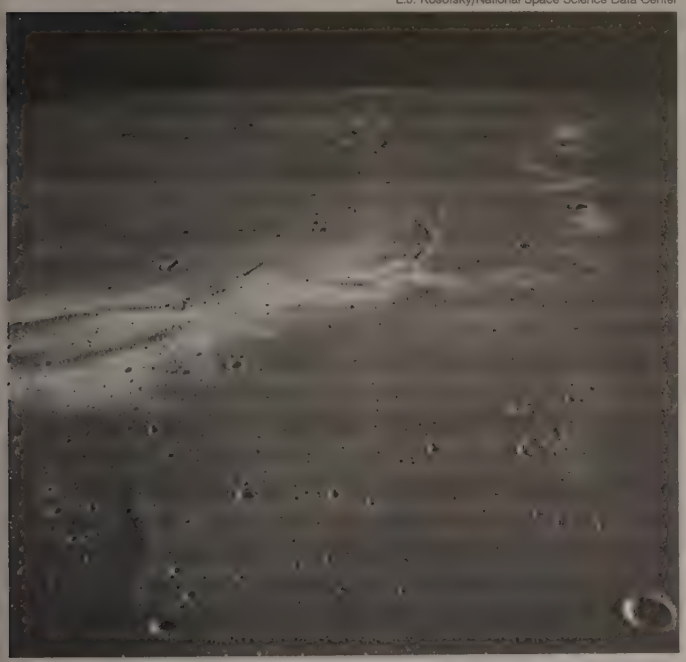


Figure 32: Reiner Gamma, photographed by Lunar Orbiter 2 in November 1966. This enigmatic lunar feature shows bright swirl patterns but no discernible topographic relief. Some scientists believe it to be the dusty trace of a comet's impact.

Among the most enigmatic features of the lunar surface are several light, swirling patterns with no associated topography. A prime example is Reiner Gamma (Figure 32), located in the southeastern portion of Oceanus Procellarum. Whereas other relatively bright features exist—*e.g.*, crater rays—they are explained as consequences of the impact process. Features such as Reiner Gamma have no clear explanation. Some scientists have suggested that they are the marks of comet impacts, in which the impacting body was large in size but had so little density as to produce no crater. Reiner Gamma is also unusual in that it coincides with a large magnetic anomaly (region of magnetic irregularity) in the crust.

Small-scale features. On a small to microscopic scale, the properties of the lunar surface are governed by a combination of phenomena—impact effects due to the arrival, at speeds up to tens of kilometres per second, of meteoritic material ranging in size down to fractions of a micrometre; bombardment by solar-wind, cosmic-ray, and solar-flare particles; ionizing radiation; and temperature extremes. Subject to no meteorological effects and unprotected by a substantial atmosphere, the uppermost surface reaches almost 400 kelvins (K; 260° F, 127° C) during the day and plunges to below 100 K (−279° F, −173° C) at night. The top layer of regolith, however, serves as an efficient insulator because of its high porosity (large number of voids, or pore spaces, per unit of volume). As a result, the daily temperature swings penetrate into the soil to less than one metre (about three feet).

Long before human beings could observe the regolith firsthand, Earth-based astronomers concluded from several kinds of measurements that the Moon's surface must be very peculiar. The evidence from photometry (brightness measurements) is particularly striking. From Earth the fully illuminated Moon is 11 times as bright as one only half illuminated, and it appears bright up to the edge of the disk. Measurements of the amount of sunlight reflected back in the direction of illumination indicate the reason: on a small scale the surface is extremely rough, and light reflected from within mineral grains and deep cavities remains shadowed until the illumination source is directly behind the observer—*i.e.*, until the full moon—at which time light abruptly reflects out of the cavities. The polarization properties of the reflected light show that the surface is rough even at a microscopic scale.

Depth of regolith

Before spacecraft landed on the Moon, astronomers had no direct means to measure the depth of the regolith layer. After the development of infrared detectors allowed them to make accurate thermal observations through the telescope, they could finally draw some reasonable conclusions about the outer surface characteristics. As Earth's shadow falls across the Moon during a lunar eclipse, the lunar surface cools rapidly, but the cooling is uneven, being slower near relatively young craters where exposed rock fields are to be expected. Some astronomers interpreted this behaviour to show that the highly insulating layer is fairly shallow, a few metres at most. This was confirmed in the mid-1960s when the first robotic spacecraft soft-landed and sank only a few centimetres instead of disappearing completely into the regolith.

Lunar rocks and soil. *General characteristics.* As noted above, the lunar regolith comprises rock fragments in a continuous distribution of particle sizes. It includes a fine fraction—dirtlike in character—that, for convenience, is called soil. The term, however, does not imply a biological contribution to its origin, as it does on Earth.

Almost all the rocks at the lunar surface are igneous—they formed from the cooling of lava. (By contrast, the most prevalent rocks exposed on Earth's surface are sedimentary, which required the action of water or wind for their formation.) The two most common kinds are basalts and anorthosites. The lunar basalts, relatively rich in iron and many also in titanium, are found in the maria. In the highlands the rocks are largely anorthosites, which are relatively rich in aluminum, calcium, and silicon. Some of the rocks in both the maria and the highlands are breccias; *i.e.*, they are composed of fragments produced by an initial impact and then reagglomerated by later impacts. The physical compositions of lunar breccias range from broken and shock-altered fragments, called clasts, to a matrix of completely impact-melted material that has lost its original mineral character. The repeated impact history of a particular rock can result in a breccia welded either into a strong, coherent mass or into a weak, crumbly mixture in which the matrix consists of poorly aggregated or metamorphosed fragments. Massive bedrock—that is, bedrock not excavated by natural processes—is absent from the lunar samples so far collected.

Lunar soils are derived from lunar rocks, but they have a distinctive character. They represent the end result of micrometeoroid bombardment and of the Moon's thermal, particulate, and radiation environments. In the ancient

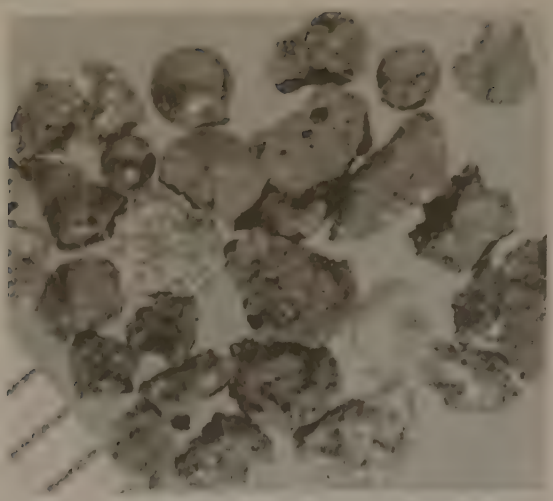


Figure 33: Discrete particles of lunar soil shown in a magnified view, part of the samples of Moon material returned by Apollo astronauts. The tiny fragments are the products of the pulverization of rocks by billions of years of meteorite and comet impacts. Major rock types represented include basalt, anorthosite, and breccia. Also present are shiny glass spherules that formed in the impacts and solidified as individual droplets. A portion of a millimetre scale is visible in the lower left corner.
NASA

Gardening of lunar soil

past the stream of impacting bodies, some of which were quite large, turned over—or "gardened"—the lunar surface to a depth that is unknown but may have been as much as tens of kilometres. As the frequency of large impacts decreased, the gardening depth became shallower. It is estimated that the top centimetre of the surface at a particular site presently has a 50 percent chance of being turned over every million years, while during the same period the top millimetre is turned over a few dozen times and the outermost tenth of a millimetre is gardened hundreds of times. One result of this process is the presence in the soil of a large fraction of glassy particles forming agglutinates, aggregates of lunar soil fragments set in a glassy cement. The agglutinate fraction is a measure of soil maturity—*i.e.*, of how long a particular sample has been exposed to the continuing rain of tiny impacts.

Although the chemical and mineralogical properties of soil particles show that they were derived from native lunar rocks, they also contain small amounts of meteoritic iron and other materials from impacting bodies. Volatile substances from comets, such as carbon compounds and water, would be expected to be mostly driven off by heat from the impact, but the small amounts of carbon found in lunar soils may include atoms of cometary origin.

A fascinating and scientifically important property of lunar soils is the implantation of solar wind particles. Unimpeded by atmospheric or electromagnetic effects, protons, electrons, and atoms arrive at speeds of hundreds of kilometres per second and are driven into the outermost surfaces of soil grains. Lunar soils thus contain a collection of material from the Sun. Because of their gardening history, soils obtained from different depths have been exposed to the solar wind at the surface at different times and therefore can reveal some aspects of ancient solar behaviour. In addition to its scientific interest, this implantation phenomenon may have implications for long-term human habitation of the Moon in the future, as discussed in the section *Lunar exploration: Lunar resources* below.

Sources of lunar material for study

The chemical and mineral properties of lunar rocks and soils hold clues to the Moon's history, and the study of lunar samples has become an extensive field of science. To date, scientists have obtained lunar material from three sources: six U.S. Apollo Moon-landing missions (1969–72), which brought back almost 382 kg (842 pounds) of samples; three Soviet Luna automated sampling missions (1970–76), which returned about 300 grams (0.66 pound) of material, and scientific expeditions to Antarctica, which have collected meteorites on the ice fields since 1969. Some of these meteorites are rocks that were blasted out of the Moon by impacts, found their way to Earth, and have been confirmed as lunar in origin by comparison with the samples returned by spacecraft. Some salient properties of the Apollo samples are listed in Table 6.

The mineral constituents of a rock reflect its chemical composition and thermal history. Rock textures—*i.e.*, the shapes and sizes of mineral grains and the nature of their interfaces—provide clues as to the conditions under which the rock cooled and solidified from a melt. The most common minerals in lunar rocks are silicates (including pyroxene, olivine, and feldspar) and oxides (including ilmenite, spinel, and a mineral discovered in rocks collected by Apollo 11 astronauts and named armalcolite, a word made from the first letters of the astronauts' surnames—Armstrong, Aldrin, and Collins). The properties of lunar minerals reflect the many differences between the history of the Moon and that of Earth. Lunar rocks appear to have formed in the total absence of water. Many minor mineral constituents in lunar rocks reflect the history of formation of the lunar mantle and crust (see below *Origin and evolution*), and they confirm the hypothesis that most rocks now found at the lunar surface formed under reducing conditions—*i.e.*, those in which oxygen was scarce.

Main groupings. The materials formed of these minerals are classified into four main groups: (1) basaltic volcanics, the rocks forming the maria, (2) pristine highland rocks uncontaminated by impact mixing, (3) breccias and impact melts, formed by impacts that disassembled and reassembled mixtures of rocks, and (4) soils, defined as unconsolidated aggregates of particles less than 1 centimetre

Table 6: Apollo Sample Summary*

	Apollo 11	Apollo 12	Apollo 14	Apollo 15	Apollo 16	Apollo 17	Apollo total
Original collection							
Number of samples	58	69	227	370	731	741	2,196
Total weight (kg)	21.6	34.3	42.3	77.3	95.7	110.5	381.7
Lithology (percent of total weight)							
Rocks (>10 mm)	44.9	80.6	67.3	74.7	72.3	65.9	69.5
Anorthosite				(0.4)	(10.5)	(0.5)	(2.8)
Basalt	(19.9)	(52.2)	(9.1)	(37.9)		(29.1)	(22.9)
Dolerite (coarse basalt)		(26.3)					(2.4)
Other igneous rocks	(2.1)	(0.2)				(4.0)	(1.3)
Breccia	(22.9)	(1.9)	(58.2)	(34.1)	(36.8)	(32.3)	(33.4)
Impact melt				(2.3)	(25.0)		(6.7)
Fines (<10 mm)	54.6	16.8	30.6	17.0	19.3	26.7	24.0
Unsieved	(54.6)	(16.8)	(0.8)	(5.1)	(8.5)	(11.3)	(11.1)
<1 mm			(27.3)	(10.7)	(9.2)	(13.9)	(11.5)
1–2 mm			(1.0)	(0.4)	(0.7)	(0.6)	(0.5)
2–4 mm			(0.8)	(0.4)	(0.5)	(0.5)	(0.4)
4–10 mm			(0.8)	(0.4)	(0.5)	(0.4)	(0.4)
Cores	0.4	1.2	0.9	6.0	7.4	6.6	5.2
Others		1.4	1.0	2.4	1.2	0.9	1.2

*Numbers in parentheses represent subtotal percentages within a group.
Source: Modified from G.H. Heiken, D.T. Vaniman, and B.M. French (eds.), *Lunar Sourcebook: A User's Guide to the Moon* (1991).

(0.4 inch) in size, derived from all the rock types. All these materials are of igneous origin, but their melting and crystallization history is complex.

Differences in lavas from Moon and Earth

The mare basalts, when in liquid form, were much less viscous than typical lavas on Earth; they flowed like heavy oil. This was due to the low availability of oxygen and the absence of water in the regions where they formed. The melting temperature of the parent rock was higher than in Earth's volcanic source regions. As the lunar lavas rose to the surface and poured out in thin layers, they filled the basins of the Moon's near side and flowed out over plains, drowning older craters and embaying the basin margins. There are also rimless craters, surrounded by dark halos, which do not have the characteristic shape of an impact scar but instead appear to have been formed by eruptions.

Most mare basalts differ from Earthly lavas not only in lacking evidence of water but also in depletion of other volatile substances such as potassium, sodium, and carbon compounds. They also are depleted of elements classified geochemically as siderophiles—elements that tend to affiliate with iron when rocks cool from a melt. (This siderophile depletion is an important clue to the history of the Earth-Moon system, as discussed in the section below *Origin and evolution*.) Some lavas were relatively rich in elements whose atoms do not readily fit into the crystal lattice sites of the common lunar minerals and are thus called incompatible elements. They tend to remain uncombined in a melt—of either mare or highland composition—and to become concentrated in the last portions to solidify upon cooling. Lunar scientists gave these lavas the name KREEP, an acronym for potassium (chemical symbol K), rare-earth elements, and phosphorus (P). These rocks give information as to the history of partial melting in the lunar mantle and the subsequent rise of lavas through the crust. Radiometric age dating (see below *Lunar exploration: Mission results*) reveals that the great eruptions that formed the maria occurred hundreds of millions of years later than the more extensive heating that produced the lunar highlands.

Ancient highland material that is considered pristine is relatively rare because most highland rocks have been subjected to repeated smashing and reagglomeration by impacts and are therefore in brecciated form. A few of the collected lunar samples, however, appear to have been essentially unaltered since they solidified in the primeval lunar crust. These rocks, some rich in aluminum and calcium or magnesium and others showing the KREEP chemical signature, suggest that late in its formation the Moon was covered by a deep magma ocean. The slow cooling of this enormous molten body, in which lighter minerals rose as they formed and heavier ones sank, appears to have resulted in the crust and mantle that exists today (see below *Origin and evolution*).

THE LUNAR INTERIOR

Structure and composition. Most of the knowledge about the lunar interior has come from the Apollo missions and from robotic spacecraft including Galileo, Clementine, and Lunar Prospector, which observed the Moon in the 1990s. Combining all available data, scientists have created a picture of the Moon (Figure 34) as a layered body comprising a low-density crust, which ranges from 60 to 100 kilometres in thickness, overlying a denser mantle, which constitutes the great majority of the Moon's volume. At the centre there probably is a small, iron-rich metallic core with a radius of about 400 kilometres at most. The core may once have been a magnetohydrodynamic dynamo like that of Earth, which could account for the remanent magnetism observed in some lunar rocks, but it appears that such internal activity has long ceased.

Asymmetry in crustal thickness

Despite these gains in knowledge, important uncertainties remain. For example, there seems to be no generally accepted explanation for the evidence that the crust is asymmetric, thicker on the Moon's far side, with the maria mostly on the near side. Examination of naturally excavated samples from large impact basins may help to resolve this and other questions in lunar history.

Internal activity of the past and present. The idea that the lunar crust is the product of differentiation in an ancient magma ocean is supported to some extent by compositional data, which show that lightweight rocks, containing such minerals as plagioclase, rose while denser materials, such as pyroxene and olivine, sank to become the source regions for the later radioactive heating episode that resulted in the outflows of mare basalts. Whether or not there ever was a uniform global ocean of molten rock, it is clear that the Moon's history is one of much heating and melting in a complex series of events that would have driven off volatiles (if any were present) and erased the record of earlier mineral compositions.

The Moon's interior

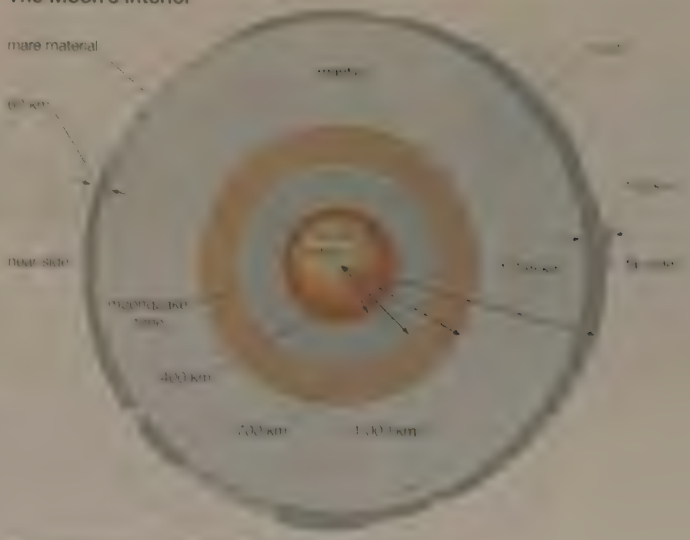


Figure 34: A cross section of the Moon's interior, showing the asymmetry in the thickness of the crust between the near and far sides and the near-side predominance of maria. Indicated distances are not to scale.

At present all evidence points to the Moon as a body in which, given its small size, all heat-driven internal processes have run down. Its heat flow near the surface, as measured at two sites by Apollo instruments, appears to be less than half that of Earth. Seismic activity is probably far less than that of Earth, though this conclusion needs to be verified by longer-running observations than Apollo provided. Many of the moonquakes detected seem to be only small "creaks" during the Moon's continual adjustment to gravity gradients in its eccentric orbit, while others are due to impacts or thermal effects. Quakes of truly tectonic origin seem to be uncommon.

Seismic activity

ORIGIN AND EVOLUTION

With the rise of scientific inquiry in the Renaissance, investigators attempted to fit theories on the origin of the Moon to the available information, and the question of the Moon's formation became a part of the attempt to explain the observed properties of the solar system (see above *An overview of the solar system: Origin of the solar system*). At first the approach was largely founded on a mathematical examination of the dynamics of the Earth-Moon system. Rigorous analysis of careful observations over a period of more than 200 years gradually revealed that, because of tidal effects, the rotations of both the Moon and Earth are slowing and the Moon is receding from Earth. Studies then turned back to consider the state of the system when the Moon was closer to Earth. Throughout the 17th, 18th, and 19th centuries, investigators examined different theories on lunar origin in an attempt to find one that would agree with the observations.

Lunar origin theories can be divided into three main categories: coaccretion, fission, and capture. Coaccretion suggests that the Moon and Earth were formed together from a primordial cloud of gas and dust. This scenario, however, cannot explain the large angular momentum of the present system. In fission theories, a fluid proto-Earth began rotating so rapidly that it flung off a mass of material that formed the Moon. Although persuasive, the theory eventually failed when examined in detail—scientists

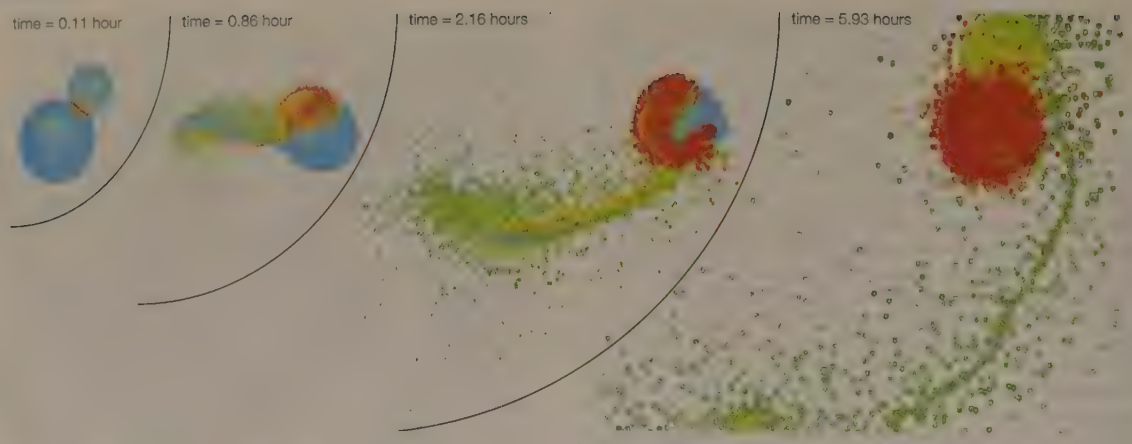


Figure 35: Stages in the first approximately six hours of a computer simulation depicting a hypothetical collision between the proto-Earth (larger object) and a Mars-sized body. Both bodies are assumed to have differentiated into an iron-rich core and an iron-depleted mantle. Matter ejected from the collision consists mainly of mantle material derived from both bodies. This material subsequently coalesces into the proto-Moon. Colours depict relative temperatures of the material heated by the collision.
From Robin M. Canup, "Simulation of a Late Lunar-Forming Impact," *Icarus*, vol. 168 (2004)

could not find a combination of properties for a spinning proto-Earth that would eject the right kind of proto-Moon. According to capture theories, the Moon formed elsewhere in the solar system and was later trapped by the strong gravitational field of Earth. This scenario remained popular for a long time, even though the circumstances needed in celestial mechanics to brake a passing Moon into just the right orbit always seemed unlikely.

By the mid-20th century, scientists had imposed additional requirements for a viable lunar origin theory. Of great importance is the observation that the Moon is much less dense than Earth, and the only likely reason is that the Moon contains significantly less iron. Such a large chemical difference argued against a common origin for the two bodies. Independent-origin theories, however, had their own problems. The question remained unresolved even after the scientifically productive Apollo missions, and it was only in the early 1980s that a model emerged—the giant-impact hypothesis—that eventually gained the support of most lunar scientists.

Giant-impact hypothesis

In this scenario (Figure 35) the proto-Earth, shortly after its formation from the solar nebula about 4.6 billion years ago, was struck a glancing blow by a body the size of Mars. Prior to the impact, both bodies already had undergone differentiation into core and mantle. The titanic collision ejected a cloud of fragments, which aggregated into a full or partial ring around Earth and then coalesced into a proto-Moon. The ejected matter consisted mainly of mantle material from the colliding body and the proto-Earth, and it experienced enormous heating from the collision. As a result, the proto-Moon that formed was highly depleted in volatiles and relatively depleted in iron (and thus also in siderophiles). Computer modeling of the collision shows that, given the right initial conditions, an orbiting cloud of debris as massive as the Moon could indeed have formed.

Once a proto-Moon was present in the debris cloud, it would have quickly swept up the remaining fragments in a tremendous bombardment. Then, over a period of 100 million years or so, the rate of impacting bodies diminished, although there still occurred occasional collisions with large objects. Perhaps this was the time of the putative magma ocean and the differentiation of the ancient plagioclase-rich crust. After the Moon had cooled and solidified enough to preserve impact scars, it began to retain the huge signatures of basin-forming collisions with asteroid-size bodies left over from the formation of the solar system. About 3.9 billion years ago one of these formed the great Imbrium Basin, or Mare Imbrium, and its mountain ramparts. During some period over the next several hundred million years occurred the long sequence of volcanic events that filled the near-side basins with mare lavas.

In an effort to unravel the history of this period, scientists have applied modern analytic techniques to lunar rock samples. The mare basalts show a wide range of chemical and mineral compositions reflecting different conditions in the deep regions of the mantle where, presumably because of heating from radioactive elements in the rock, primordial lunar materials were partly remelted and fractionated so that the lavas carried unique trace-element signatures up to the surface. By studying the past events and processes reflected in the mineral, chemical, and isotopic properties of these rocks, scientists have slowly built a picture of a variegated Moon. Their findings have provided valuable background information for Earth- and spacecraft-based efforts to map how the content of important materials varies over the lunar surface.

Second heating episode

Once the huge mare lava outflows had diminished, apparently the Moon's heat source had run down. The last few billion years of its history have been calm and essentially geologically inactive except for the continuing rain of impacts, which is also declining over time, and the microscopic weathering due to bombardment by solar and cosmic radiation and particles.

LUNAR EXPLORATION

Early studies. Investigations of the Moon and some understanding of lunar phenomena can be traced back to a few centuries before the Christian era. Knowledge about the Moon accumulated slowly at first, driven by astrological and navigational needs, until an outburst of progress began in the Renaissance. In the early 1600s the German astronomer Johannes Kepler used observations made by Tycho Brahe of Denmark to find empirically the laws governing planetary motion. In 1609–10 Galileo began his telescopic observations that forever changed human understanding of the Moon. Most effort hitherto had been devoted to understanding the movements of the Moon through space, but now astronomers began to focus their attention on the character of the Moon as a world of its own. Some milestones in human exploration and understanding of the Moon are given in Table 7.

Exploration by spacecraft. *First robotic missions.* Following the launch in 1957 of the U.S.S.R.'s satellite Sputnik, the first spacecraft to orbit Earth, it became obvious that the next major goal of both the Soviet and the U.S. space programs would be the Moon. The United States quickly launched a few robotic lunar probes, most of which failed and none of which reached the Moon. The Soviet Union had more success, achieving in 1959 the first escape from Earth's gravity with Luna 1, the first impact on the lunar surface with Luna 2, and the first photographic survey of the Moon's far side with Luna 3.

Luna missions

After the National Aeronautics and Space Administration (NASA) was founded in 1958, the U.S. program became more ambitious technically and more scientifically orient-

Table 7: History of Lunar Observation and Exploration

time period	accomplishment
Prehistoric and early historic times	Basic knowledge of the Moon's motion, phases, and markings reflected in legends.
500 BC to AD 150	Phases and eclipses correctly explained; Moon's size and its distance from Earth measured.
Middle Ages	Lunar ephemeris refined.
Renaissance	Laws of motion formulated; telescopic observations begun.
19th century	Near-side mapping completed; atmosphere proved absent; geologic principles applied in volcanism-versus-impact debate.
1924	Polarimetry used to show that lunar surface is composed of small particles.
1927–30	Surface temperatures measured for lunar day and night and during eclipses.
1946	Radar echoes reflected from Moon and detected for first time.
1950–57	Theories of Moon's formation incorporated in efforts to explain origin of solar system; radiometric age dating employed in meteorite research; lunar subsurface temperatures measured by microwave radiometry; relative ages of lunar features derived from principles of stratigraphy.
1959	Luna 2 spacecraft first man-made object to strike Moon; global magnetic field found to be absent; Luna 3 supplies first far-side images.
1960	Detailed measurements of lunar surface cooling during eclipses made from Earth.
1964	Ranger 7 transmits high-resolution pictures of Moon.
1966	Luna 9 and Surveyor 1 make first lunar soft landings; Luna 10 and Lunar Orbiter 1 first spacecraft to orbit Moon.
1967	First measurements made of lunar surface chemistry.
1968	Mascons discovered in analysis of data from Lunar Orbiters; Apollo 8 astronauts orbit Moon.
1969	Apollo 11 astronauts first humans to walk on Moon; lunar samples and data returned to Earth.
1969–74	Manned Apollo orbital and surface expeditions and automated Luna flights explore Moon's lower latitudes; Apollo program completed.
1970s–present	Lunar studies continued using samples returned by Apollo and Luna missions, meteorites originating from Moon, and data gathered by Earth-based mineralogical remote-sensing techniques.
1990	Galileo spacecraft collects compositional remote-sensing data during lunar flyby, demonstrating potential for future orbital geochemical missions.
1994	Orbiting Clementine spacecraft provides imagery, altimetry, and gravity maps of entire Moon.
1998–99	Orbiting Lunar Prospector spacecraft maps lunar surface composition and magnetic field; its neutron spectrometer data suggests presence of water ice at both poles.

ed. Initial spacecraft investigations were geared toward studying the Moon's fundamental character as a planetary body by means of seismic observation, gamma-ray spectrometry, and close-up imaging. Scientists believed that even limited seismic data would give clues toward resolving the question of whether the Moon was a primitive, undifferentiated body or one that had been heated and modified by physical and chemical processes such as those on Earth. Gamma-ray measurements would complement the seismic results by showing whether the Moon's interior had sufficient radioactivity to serve as an active heat engine, and they would also generate data about the chemical composition of the lunar surface. Imaging would reveal features too small to be seen from Earth, perhaps providing information on lunar surface processes and also arousing public interest.

Among nine U.S. Ranger missions launched between 1961 and 1965, Ranger 4 (1962) became the first U.S. spacecraft to strike the Moon. Only the last three craft, however, avoided the plaguing malfunctions that limited or prematurely ended the missions of their predecessors. Ranger 7 (1964) returned thousands of television images before impacting as designed, and Rangers 8 and 9 (both 1965) successfully followed. The impact locale of Ranger 7 was named Mare Cognitum for the new knowledge gained, a major result being the discovery that even small lunar features have been mostly subdued from incessant meteorite impacts.

After a number of failures in the mid-1960s, the Soviet Union scored several notable achievements: the first successful lunar soft landing by Luna 9 and the first lunar orbit by Luna 10, both in 1966. Pictures from Luna 9 revealed the soft, rubbly nature of the regolith and, because the landing capsule did not sink out of sight, confirmed its approximate bearing strength. Gamma-ray data from Luna 10 hinted at a basaltic composition for near-side regions. In 1965 the Soviet flyby mission designated Zond 3 returned good pictures of the Moon's far side.

In the mid-1960s the United States carried out its own soft-landing and orbital missions. In 1966 Surveyor 1 touched down on the Moon and returned panoramic television images. Six more Surveyors followed between 1966 and 1968, with two failures; they provided not only detailed television views of lunar scenery but also the first chemical data on lunar soil and the first soil-mechanics information showing mechanical properties of the top few centimetres of the regolith. Also, during 1966–67, five U.S. Lunar Orbiters made photographic surveys of most of the lunar surface, providing the mapping essential for planning the Apollo missions.

Surveyor television images

Apollo and later missions. After the Soviet cosmonaut Yury Gagarin pioneered human Earth-orbital flight in April 1961, U.S. President John F. Kennedy established the national objective of landing a man on the Moon and returning him safely by the end of the decade. Apollo was the result of that effort.

Within a few years the Soviet Union and the United States were heavily engaged in a political and technological race to launch manned flights to the Moon. At the time, the Soviets did not publicly acknowledge the full extent of their program, but they did launch a number of human-precursor circumlunar missions between 1968 and 1970 under the generic name Zond, using spacecraft derived from their piloted Soyuz design. Some of the Zond flights brought back colour photographs of the Moon's far side and safely carried live tortoises and other organisms around the Moon and back to Earth. In parallel with these developments, Soviet scientists began launching a series of robotic Luna spacecraft designed to go into lunar orbit and then land with heavy payloads. This series, continuing to 1976, eventually returned drill-core samples of regolith to Earth and also landed two wheeled rovers, Lunokhod 1 and 2 (1970 and 1973), that pioneered robotic mobile exploration of the Moon.

In December 1968, acting partly out of concern that the Soviet Union might be first in getting people to the Moon's vicinity, the United States employed the Apollo 8 mission to take three astronauts—Frank Borman, James Lovell,

and William Anders—into lunar orbit. After circling the Moon three times, the crew returned home safely with hundreds of photographs. The Apollo 9 and 10 missions completed the remaining tests of the systems needed for landing on and ascending from the Moon. On July 20, 1969, Apollo 11 astronauts Neil Armstrong and Edwin ("Buzz") Aldrin set foot on the Moon, while Michael Collins orbited above them. Five successful manned landing missions followed, ending with Apollo 17 in 1972; at the completion of the program, a total of 12 astronauts had set foot on the Moon.

First manned landing

Twenty years later the Soviet Union admitted that it had indeed been aiming at the same goal as Apollo, not only with a set of spacecraft modules for landing on and returning from the Moon but also with the development of a huge launch vehicle, called the N1, comparable to the Apollo program's Saturn V. After several launch failures of the N1, the program was canceled in 1974. (For a comprehensive description of the U.S. and Soviet Moon programs, see EXPLORATION: *Space exploration: From Sputnik to Apollo: The race to the Moon.*)

After the Apollo missions, lunar scientists continued to conduct multispectral remote-sensing observations from Earth and perfected instrumental and data-analysis techniques. During Galileo's flybys of Earth and the Moon in December 1990 and 1992 en route to Jupiter, the spacecraft demonstrated the potential for spaceborne multispectral observations—*i.e.*, imaging the Moon in several discrete wavelength ranges—to gather geochemical data. As a next logical step, scientists generally agreed on a global survey by an automated spacecraft in polar orbit above the Moon and employing techniques evolved from those used during the Apollo missions. After a long hiatus, orbital mapping of the Moon resumed with the flights of the Clementine and Lunar Prospector spacecraft, launched in 1994 and 1998, respectively.

Mission results. The Apollo program revolutionized human understanding of the Moon. The samples collected and the human and instrumental observations have continued to be studied into the 21st century. Analyses of samples from the Luna missions have continued as well and are valuable because they were collected from eastern equatorial areas far from the Apollo sites.

Radiometric dating of samples

One new and fundamental result has come from radiometric age dating of the samples. When a rock cools from the molten to the solid state, its radioactive isotopes are immobilized in mineral crystal lattices and then decay in place. Knowing the rate of decay of one nuclear species (nuclide) into another, scientists can, in principle, use the ratios of decay products as a clock to measure the time elapsed since the rock cooled. Some nuclides, such as isotopes of rubidium and strontium, can be used to date rocks billions of years old. With great care in sample preparation and use of mass spectrometry, the isotopic ratios can be found and converted into age estimates. By the time of the Apollo sample returns, scientists had refined this art, and, using meteorite samples, they were already investigating the early history of the solar system.

Analysis of the first lunar samples confirmed that the Moon is an evolved body with a long history of differentiation and volcanic activity. Unlike the crust of Earth, however, the lunar crust is not recycled by tectonic processes, so it has preserved the records of ancient events. Highland rock samples returned by the later Apollo missions are nearly four billion years old, revealing that the Moon's crust was already solid soon after the planets condensed out of the solar nebula. The mare basalts, though they cover a wide range of ages, generally show that the basin-filling volcanic outpourings occurred long after the formation of the highlands; this is the reason they are believed to have originated from later radioactive heating within the Moon rather than during the primordial heating event. Trace-element analyses indicate that the magmatic processes of partial melting gave rise to different lavas.

In addition to collecting samples, Apollo astronauts made geologic observations, took photographs, and placed long-lived instrument arrays and retroreflectors on the lunar surface. Not only the landing expeditions but also the Apollo orbital observations yielded important new knowledge. On each mission the Moon-orbiting Command and Service modules carried cameras and remote-sensing instruments for gathering compositional information.

Results of Clementine and Lunar Prospector

The Clementine and Lunar Prospector spacecraft, operating in lunar polar orbits, used complementary suites of remote-sensing instruments to map the entire Moon, measuring its surface composition, geomorphology, topography, and gravitational and magnetic anomalies (see Figure 29 and Figure 36). The topographic data highlighted the huge South Pole–Aitken Basin, which, like the other basins on the far side, is devoid of lava filling. Measuring roughly 2,500 kilometres in diameter and 13 kilometres deep, it is the largest impact feature on the Moon and the largest known in the solar system; because of its location, its existence was not confirmed until the Lunar Orbiter missions in the 1960s. The gravity data collected by the spacecraft, combined with topography, confirmed the existence of a thick rigid crust, giving yet more evidence that the Moon's heat source has expired. Both spacecraft missions hinted at the long-considered possibility that water ice exists in permanently shadowed polar craters. The most persuasive evidence came from the neutron spectrometer of Lunar Prospector (see below).

Lunar resources. Scientists and space planners have long acknowledged that extended human residence on the Moon would be greatly aided by the use of local resources. This would avoid the high cost of lifting payloads against Earth's strong gravity. Certainly, lunar soil could be used for shielding habitats against the radiation environment. More advanced uses of lunar resources are clearly possible, but how advantageous they would be is presently un-

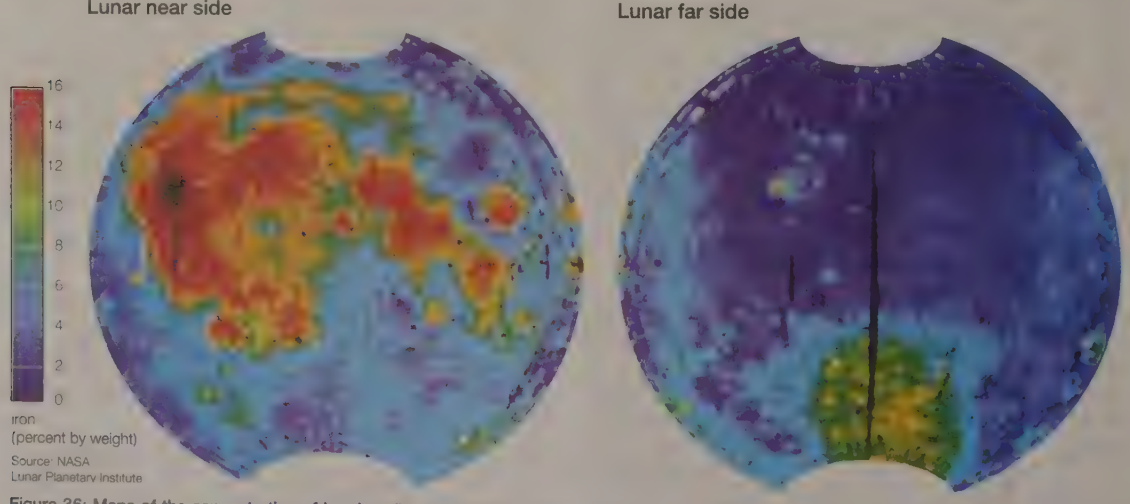


Figure 36: Maps of the concentration of iron in soils on the Moon's near and far sides, derived from multispectral imaging data collected by the Clementine spacecraft in 1994.

known. For example, most lunar rocks are about 40 percent oxygen, and chemical and electrochemical methods for extracting it have been demonstrated in laboratories. Nevertheless, significant engineering advances would be needed before the cost and difficulty of operating an industrial-scale mining and oxygen-production facility on the Moon could be estimated and its advantages over transporting oxygen from Earth could be evaluated. In the long run, however, some form of extractive industry on the Moon is likely, in part because launching fleets of large rockets continuously from Earth would be too costly and too polluting of the atmosphere.

The solar wind has implanted hydrogen, helium, and other elements in the surfaces of fine grains of lunar soil. Though their amounts are small, about 100 parts per million in the soil, they may someday serve as a resource. They are easily released by moderate heating, but large volumes of soil would need to be processed to obtain useful amounts of the desired materials. Helium-3, a helium isotope that is rare on Earth and that has been deposited on the Moon by the solar wind, has been proposed as a fuel for nuclear fusion reactors in the future.

One natural resource uniquely available on the Moon is its polar environment. Because the Moon's axis is nearly perpendicular to the plane of the ecliptic, sunlight is always horizontal at the lunar poles, and certain areas, such as crater bottoms, exist in perpetual shadow. Under these conditions the surface may reach temperatures as low as 40 K (−388° F, −233° C). Some scientists have theorized that these cold traps may have collected volatile substances, including water ice, over geologic time, though others have expressed doubt that ice deposits could have survived there.

Lunar Prospector, which orbited the Moon for a year and a half, carried a neutron spectrometer to investigate the composition of the regolith within about a metre of the surface. Neutrons originating underground due to radioactivity and cosmic-ray bombardment interact with the nuclei of elements in the regolith en route to space, where they can be detected from orbit. A neutron loses more energy in an interaction with a light nucleus than with a heavy one, so the observed neutron spectrum can reveal whether light elements are present in the regolith. Lunar Prospector gave clear indications of light-element concentrations at both poles, interpreted as proof of excess hydrogen atoms. The observed hydrogen signature may represent the theoretically predicted deposits of water ice.

Suggestions of polar ice

A high priority for future lunar exploration is to send an autonomous robotic rover into a dark polar region to confirm the putative ice deposits, find out the form of the ice if it exists, and begin assessing its possible utility. If lunar ice can be mined economically, it can serve as a source of rocket propellants when split into its hydrogen and oxygen components. From a longer-term perspective, however, the ice would better be regarded as a limited, recyclable resource for life support (in the form of drinking water and perhaps breathable oxygen). Should this resource exist, an international policy for its conservation and management will be needed.

Even if no icy bonanza is discovered, the lunar polar regions still represent an important resource. Only there can be found not only continuous darkness but also continuous sunlight. A solar collector tracking the Sun from a high peak near a lunar pole could provide essentially uninterrupted heat and electric power. The lunar poles also could serve as good sites for certain astronomical observations. To observe objects in the cosmos that radiate in the infrared and millimetre-wavelength regions of the spectrum, astronomers need telescopes and detectors that are cold enough to limit the interference generated by the instruments' own heat. To date, such telescopes launched into space have carried cryogenic coolants, which eventually run out. A telescope permanently sited in a lunar polar cold region and insulated from local heat sources might cool on its own to 40 K (−388° F, −233° C) or lower. Although such an instrument could observe less than half the sky—ideally, one would be placed at each lunar pole—it would enable uninterrupted viewing of any object above its horizon. (J.D.Bu.)

Mars

Mars, designated ♂ in astronomy, is the fourth planet in the solar system in order of distance from the Sun and seventh in size and mass. It is a conspicuous, sometimes quite bright, reddish object in the night sky. Because of its blood-red colour—it is sometimes called the Red Planet—Mars has long been associated with warfare and slaughter. It is named for the Roman god of war. The planet's two moons, Phobos (Greek: "Fear") and Deimos ("Terror"), were named for two of the sons of Ares (the counterpart of Mars in Greek mythology) and Aphrodite.

Mars is the second closest planet to Earth, after Venus, and it is usually easy to observe in the night sky because its orbit lies outside Earth's. It is also the only planet whose solid surface and atmospheric phenomena can be seen in telescopes from Earth. Like Earth, Mars has clouds, winds, a roughly 24-hour day, seasonal weather patterns, polar caps, volcanoes, canyons, and other familiar features.

At the end of the 19th century, speculation was rife that the so-called canals of Mars—complex systems of long, straight surface lines that a very few astronomers had claimed to see in telescopic observations—were the creations of intelligent beings. Although the canals later proved to be illusory, scientific and public interest in the possibility of Martian life and in exploration of the planet has not faded. Indeed, there are intriguing clues that billions of years ago Mars was even more Earth-like than today, with a denser, warmer atmosphere and much more water—rivers, lakes, flood channels, and perhaps oceans. By all indications, Mars is now a sterile frozen desert, but close-up images from the Mars Global Surveyor spacecraft of seemingly water-eroded gullies suggest that at least small amounts of water may have flowed on or near the planet's surface in geologically recent times and may still exist as a liquid in protected areas below the surface. If microscopic life-forms ever did originate on Mars, there remains a chance that they may yet survive in these watery niches.

Clues to a more Earth-like past

BASIC ASTRONOMICAL DATA

Mars moves around the Sun at a mean distance of 228 million kilometres, or about 1.5 times that of Earth from the

NASA/JPL/Malin Space Science Systems

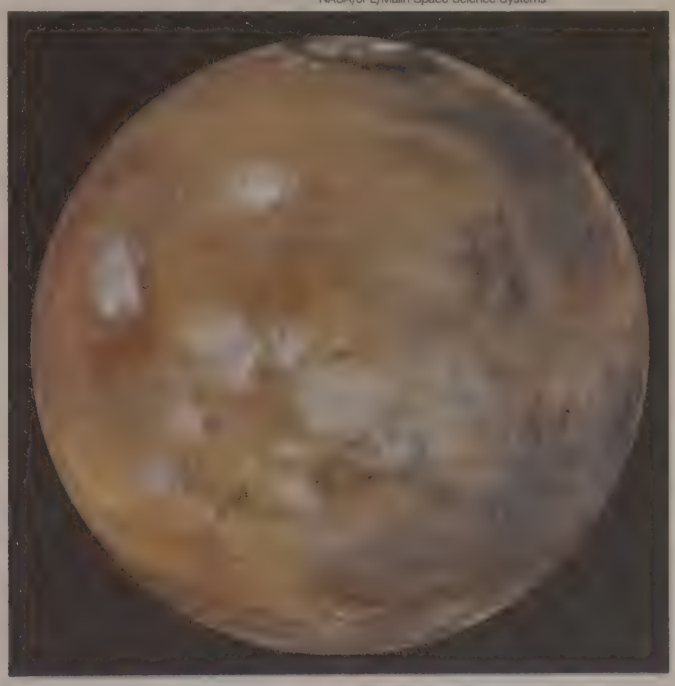


Figure 37: Global image of Mars constructed from 24 photographs taken on a day in April 1999 by the Mars Global Surveyor spacecraft. Wispy water-ice clouds hang over the line of Tharsis volcanoes (west of centre) and Olympus Mons and Alba Patera (northwest and north of Tharsis, respectively). East of Tharsis lies the Valles Marineris canyon system. The north polar cap is at the top of the image.

Sun. Owing to its relatively elongated orbit, the distance between Mars and the Sun varies from 206.6 million to 249.2 million kilometres. Mars orbits the Sun once in 687 Earth days, which means that its year is nearly twice as long as Earth's. At its closest approach, Mars is less than 56 million kilometres from Earth, but it recedes to almost 400 million kilometres when the two planets are on opposite sides of the solar system.

Mars is easiest to observe when it is in the opposite direction in the sky to the Sun—*i.e.*, at opposition—because it is then high in the sky and shows a fully lighted face. Successive oppositions occur about every 26 months. Oppositions can take place at different points in the Martian orbit. Those best for viewing occur when the planet is closest to the Sun, and so also to Earth, because Mars is then at its brightest and largest. Close oppositions occur roughly every 15 years.

Table 8: Orbital Data for Mars

Mean distance from Sun	227,941,040 km (1.5 AU)
Eccentricity of orbit	0.093399
Inclination of orbit to ecliptic	1.85020°
Martian year (sidereal period of revolution)	686.9800 Earth days
Mean synodic period	779.94 Earth days
Mean orbital velocity	24.1 km/s

Mars spins on its axis once every 24 hours 37 minutes, making a day on Mars only a little longer than an Earth day. Its axis of rotation is inclined to its orbital plane by about 25°, and, as for Earth, the tilt gives rise to seasons on Mars. The Martian year consists of 668.6 Martian solar days, called sols. At present, southern summer occurs when Mars is closest to the Sun in its orbit. As a result, southern summers are shorter (154 Martian days) and warmer than those in the north (178 Martian days). The situation, however, is slowly changing such that 25,000 years from now the northern summers will be the shorter and warmer ones. In addition, the tilt of the axis is slowly changing on a roughly one-million-year timescale. During some epochs the tilt is close to zero, and so Mars has no seasons; in other epochs the tilt may be as high as 35°, resulting in extreme seasonal differences.

Martian seasons

Table 9: Physical Data for Mars

Equatorial radius	3,396.2 km
North polar radius	3,376.2 km
South polar radius	3,382.6 km
Surface area	1.44×10^8 km^2
Mass	6.418×10^{23} kg
Mean surface gravity	372 cm/s^2
Mean density	3.94 g/cm^3
Escape velocity	5.022 km/s
Rotation period (Martian sidereal day)	24 h 37 min 22.663 s
Martian mean solar day (sol)	24 h 39 min 36 s
Inclination of equator to orbit	24.936°
Mean surface temperature	210 K (−82° F, −63° C)
Typical surface pressure	0.006 bar
Number of known moons	2

Mars is a small planet, larger only than Pluto and Mercury and slightly more than half the size of Earth. It has an equatorial radius of 3,396 kilometres and a mean polar radius of 3,379 kilometres, both values accurately determined by the Mars Global Surveyor spacecraft, which began its primary mission in orbit around the planet in 1999. The mass of Mars is only one-tenth the terrestrial value, and its gravitational acceleration of 3.72 metres (12.2 feet) per second per second at the surface means that objects on Mars weigh a little more than a third of that on Earth's surface. Mars has only 28 percent of the surface area of Earth, but, because about three-fourths of Earth is covered by water, the land areas of the two planets are comparable. For additional orbital and physical data, see Tables 8 and 9.

EARLY TELESCOPIC OBSERVATIONS

Mars was an enigma to ancient astronomers, who were bewildered by its apparently capricious motion across the sky—sometimes in the same direction as the Sun and other celestial objects (direct, or prograde, motion), sometimes in the opposite direction (retrograde motion). In 1609 the German astronomer Johannes Kepler used the superior naked-eye observations of the planet by his Danish colleague Tycho Brahe to deduce empirically its laws of motion and so pave the way for the modern gravitational theory of the solar system. Kepler found that the orbit of Mars was an ellipse along which the planet moved with nonuniform but predictable motion. Earlier astronomers had based their theories on the older Ptolemaic idea of hierarchies of circular orbits and uniform motion.

The earliest telescopic observations of Mars in which the disk of the planet was seen were those of the Italian astronomer Galileo in 1610. The Dutch scientist and mathematician Christiaan Huygens is credited with the first accurate drawings of surface markings. In 1659 Huygens made a drawing of Mars showing a major dark marking on the planet now known as Syrtis Major. The Martian polar caps were first noted by the Italian-born French astronomer Gian Domenico Cassini about 1666.

Visual observers made many key discoveries. The rotation of the planet was discovered by Huygens in 1659 and measured by Cassini in 1666 to be 24 hours 40 minutes—in error by only 3 minutes. The tenuous Martian atmosphere was first noted in the 1780s by the German-born British astronomer William Herschel, who also measured the tilt of the planet's rotation axis and first discussed the seasons of Mars. In 1877 Asaph Hall of the United States Naval Observatory discovered that Mars has two natural satellites. Visual observations also documented many meteorological and seasonal phenomena that occur on Mars, such as various cloud types, the growing and shrinking of the polar caps, seasonal changes in the colour and extent of the dark areas, an annual "wave of darkening" in the markings that sweeps across the planet in phase with the shrinking of the polar caps, and an occasional "blue haze" in the atmosphere. The explanation of most of these, however, had to await the exploration of Mars by spacecraft.

Discovery of Mars's moons

GENERAL APPEARANCE

To the Earth-based telescopic observer, the Martian surface outside the polar caps is characterized by red-ochre-coloured bright areas on which dark markings appear superposed. In the past, the bright areas were referred to as deserts, and the majority of large dark areas were originally called *maria* (Latin: "oceans" or "seas"; singular *mare*) in the belief that they were covered by expanses of water.

Surface features. The dark markings cover about one-third of the Martian surface, mostly in a band around the planet between latitudes 10° and 40° S. Their distribution is irregular, and their gross pattern has been observed to change over timescales of tens to hundreds of years. The northern hemisphere has only three such major features—Acidalia Planitia, Syrtis Major, and a dark collar around the pole—which were once considered to be shallow seas or vegetated regions. It is now known that Mars's dark areas form and change as winds move materials around the surface. Viewed in close-up images taken by spacecraft, they are seen to be formed by many separate and grouped dark streaks and splotches that are associated with craters, ridges, hills, and other obstructions to the flow of local winds. Seasonal and longer-term variations in size and colour of the dark areas probably reflect changes in the proportion of surface covered by various materials.

The bright areas, which represent about two-thirds of the planet's surface, display subtle shadings, but these probably also are the result of differences in the distribution of wind-blown materials on the surface. The canals that figured so prominently on maps made from telescopic observations around the turn of the 20th century (see below *The mapping of Mars*) are not visible in close-up spacecraft images from flyby and orbital missions, beginning with Mariner 4 in 1965. They were almost certainly imaginary features that observers thought they saw while straining to make out objects close to the limit of resolution of their telescopes. Other features, such as the "wave of darkening" and the "blue haze" described by early observers at the telescope, are now known to result from a combination of the

Absence of canals in spacecraft images

viewing conditions and changes in the reflective properties of the surface.

Polar regions. For telescopic observers, the most striking regular changes on Mars occur at the poles. With the onset of fall in a particular hemisphere, clouds develop over the relevant polar region, and the cap, made of frozen carbon dioxide, begins to grow. The smaller cap in the north ultimately extends to 55° latitude; the larger one in the south to 50° latitude. In spring the caps recede. During summer the northern carbon dioxide cap disappears completely, leaving behind a small water-ice cap. In the south a small residual carbon dioxide cap lingers over the summer; no water-ice cap has been detected there (although one may lie beneath the carbon dioxide cap).

Transient atmospheric phenomena. Early telescopic observers noted instances in which Martian surface features were temporarily obscured. They generally classified the cause as being white clouds or yellow clouds, which were correctly interpreted as due to condensed gas or dust, respectively. Spacecraft observations have confirmed that hazes, clouds, and fogs commonly veil the surface.

Images from spacecraft in Mars orbit have documented a variety of low-lying clouds and fogs that are often confined within depressions—*i.e.*, valleys or craters. They have also revealed high, thin clouds, particularly at the morning terminator (the dividing line between the lit and unlit portions of the planet's disk). Orographic clouds, produced when moist air is lifted over elevated terrain and cooled, form around prominent features such as craters and volcanoes, and winter at mid-latitudes is characterized by westward-moving, spiral-shaped storm systems, similar to those on Earth. Most of these clouds are composed of water ice—the white clouds seen by the early observers.

Dust storms are common on Mars. They can occur at any time but are most frequent in southern spring and summer, when Mars is passing closest to the Sun and surface temperatures are at their highest. Most of the storms are regional in extent and last a few weeks. Every second or third year, however, the dust storms become global. At their peak, dust is carried so high in the atmosphere that only the summits of the loftiest volcanoes—up to 21 kilometres above the planet's mean radius—are visible.

Dust devils

Although too small to be observed from Earth, dust devils have been seen from Mars orbit and at the landing site of the Mars Pathfinder spacecraft. Narrow tracks, thought to be caused by dust devils, are also visible in high-resolution images taken from orbit by Mars Global Surveyor.

THE ATMOSPHERE

Basic atmospheric data. The American astronomer Gerard P. Kuiper ascertained from telescopic observations in 1947 that the Martian atmosphere is composed mainly of carbon dioxide. The atmosphere is very thin, exerting less than 1 percent of Earth's atmospheric pressure at the surface. Surface pressures range over a factor of 15 owing to the large altitude variations in Mars's topography. Only small amounts of water are present in the atmosphere today. If it all precipitated out, it would form a layer of ice crystals only 10 micrometres (0.0004 inch) thick, which could be gathered into a solid block of ice not much larger than a medium-size terrestrial iceberg. As discussed below, geologic evidence suggests that the atmosphere was much denser in the remote past and that water was once much more abundant at the surface.

Temperatures in the lower atmosphere

The characteristic temperature in the lower atmosphere is about 200 kelvins (K; −100° F, −70° C), which is generally colder than the average daytime surface temperature of 250 K (−10° F, −20° C). These values are in the same range as those experienced on Earth in Antarctica during winter. In summer above a very dark surface, daytime temperatures at the height of a human being can peak at about 290 K (62° F, 17° C). Above the turbulent layer close to the surface, temperature decreases with elevation at a rate of about 1.5 K (2.7° F, 1.5° C) per kilometre of altitude.

Unlike that of Earth, the atmosphere of Mars experiences large seasonal variations in pressure as carbon dioxide, the main constituent, "snows out" at the winter pole and returns directly to a gas (sublimes) in the spring. Because the southern winter cap is more extensive than the northern, atmospheric pressure reaches a minimum during southern winter when the southern cap is at its largest. From Viking lander measurements of the pressure, which was found to vary by 26 percent annually over the mean, scientists calculated that some 7.9 trillion metric tons of carbon dioxide leave and reenter the atmosphere seasonally. This is equivalent to a thickness of at least 23 centimetres (9 inches) of solid carbon dioxide (dry ice) or several metres of carbon dioxide snow averaged over the vast area of the seasonal polar caps.

Composition and surface pressure. Direct chemical analysis at the surface by the Viking landers and spectral observations from orbiting spacecraft have allowed scientists to determine the composition of the Martian atmosphere to high precision. Where the atmosphere is well mixed by turbulence—below an altitude of 125 kilometres—95.3 percent of the atmosphere by weight is carbon dioxide (see Table 10). This is a comparatively large amount—nine times the quantity now in Earth's much more massive atmosphere. Much of Earth's carbon dioxide, however, is chemically locked in sedimentary rocks; the amount in the Martian atmosphere is less than a thousandth of the terrestrial total. The balance of the Martian atmosphere consists of molecular nitrogen, water vapour, and noble gases (argon, neon, krypton, and xenon). There are also trace amounts of gases that have been produced from the primary constituents by photochemical reactions, generally high in the atmosphere; these include molecular oxygen, carbon monoxide, nitric oxide, and small amounts of ozone.

The lower atmosphere supplies gas to the planet's ionosphere, where densities are low, temperatures are high, and components separate by diffusion according to their masses. Various constituents in the top of the atmosphere are lost to space, which affects the isotopic composition of the remaining gases. For example, because hydrogen is lost preferentially over its heavier isotope deuterium, Mars's atmosphere contains five times more deuterium than Earth's.

Table 10: Composition of the Martian Atmosphere
(percentage by weight)

Carbon dioxide (CO_2)	95.32
Molecular nitrogen (N_2)	2.7
Argon (Ar)	1.6
Molecular oxygen (O_2)	0.13
Carbon monoxide (CO)	0.07
Water vapour (H_2O)	0.03
Neon (Ne)	0.00025
Krypton (Kr)	0.00003
Xenon (Xe)	0.000008

Although water is only a minor constituent of the Martian atmosphere (a few molecules per 10,000 at most), primarily because of low atmospheric and surface temperatures, it plays an important role in atmospheric chemistry and meteorology. The Martian atmosphere is effectively saturated with water vapour, yet there is no liquid water present on the surface. The temperature and pressure of the planet are so low that water can exist only as ice or as vapour. Little water is exchanged daily with the surface despite the very cold nighttime surface temperatures.

Subsurface ice

The atmospheric water vapour is believed to be in contact with a much larger reservoir in the Martian soil. Layers of ice just below the surface are thought to be ubiquitous on Mars at latitudes poleward of 40°; the very low subsurface temperatures would prevent the ice from subliming. The Mars Odyssey spacecraft, which began orbital observations of the planet in late 2001, confirmed that ice is present within a metre of the surface at these latitudes, but it is not known how deep the ice layer extends. In contrast, at low latitudes, any ice present in the ground would tend to sublime into the atmosphere.

Most of the information about atmospheric water on Mars has come from the Viking orbiters, which observed seasonal patterns of water content in the atmosphere over a full Martian year. The Viking landers' analytical instruments also measured the isotopic composition of the major and minor atmospheric gases. The similarity of the ratios

of isotopes of carbon and oxygen to their terrestrial values implies that large reservoirs of carbon dioxide and water ice exist on Mars and that gases from the reservoirs exchange with those in the atmosphere. The results of other isotopic measurements from Viking suggest that larger amounts of carbon dioxide, nitrogen, and argon were present in the atmosphere in the past and that Mars may have lost much of its inventory of volatile substances early in its history, either to space or to the ground (*i.e.*, locked up chemically in rocks). Some scientists have conjectured on the basis of these ideas that Mars may once have had a much thicker atmosphere.

Atmospheric structure. The vertical structure of the Martian atmosphere—that is, the relation of temperature and pressure to altitude—is determined partly by a complicated balance of several energy transport mechanisms and partly by the way energy from the Sun is introduced into the atmosphere and lost by radiation to space.

Two factors control the vertical structure of the lower atmosphere—its composition of almost pure carbon dioxide and its content of large quantities of suspended dust. Because carbon dioxide radiates energy efficiently at Martian temperatures, the atmosphere can respond rapidly to changes in the amount of solar radiation received. The suspended dust absorbs large quantities of heat directly from sunlight and provides a distributed source of energy throughout the lower atmosphere.

Daily temperature swings

Surface temperatures depend on latitude and fluctuate over a wide range from day to night. At the Viking 1 and Pathfinder landing sites (both about 20° N latitude), the temperatures at roughly human height above the surface regularly varied from a low near 189 K (−119° F, −84° C) just before sunrise to a high of 240 K (−28° F, −33° C) in the early afternoon. This temperature swing is much larger than occurs in desert regions on Earth. The variation is greatest very close to the ground and occurs because the thin dry atmosphere allows the surface to radiate its heat quickly during the night. During dust storms this ability is impaired, and the temperature swing is reduced. Above altitudes of a few kilometres, the daily variation is damped out, but other oscillations appear throughout the atmosphere as a result of the direct input of solar energy. These temperature and pressure oscillations give the Martian atmosphere a very complex vertical structure.

The cooling of the atmosphere with altitude at a rate of 1.5 K per kilometre continues upward to about 40 kilometres, at which level (called the tropopause) the temperature becomes a roughly constant 140 K (−210° F, −130° C). This rate, measured by the Viking (and later Pathfinder) spacecraft as they descended through the atmosphere, was unexpectedly low; scientists had anticipated it to be near 5 K per kilometre. Moreover, the tropopause was expected to occur at a much lower altitude, about 15 kilometres. The large amount of suspended dust is thought to be responsible for these differences.

Above 100 kilometres, the structure of the atmosphere is determined by the tendency of the heavier molecules to concentrate below the lighter ones. This diffusive separation process overcomes the tendency of turbulence to mix all the constituents together. At these high altitudes, absorption of ultraviolet light from the Sun dissociates and ionizes the gases and leads to complex sequences of chemical reactions. The top of the atmosphere has an average temperature of about 300 K (80° F, 27° C).

Meteorology and atmospheric dynamics. The global pattern of atmospheric circulation on Mars shows many superficial similarities to that of Earth, but the root causes are very different. Among these differences are the atmosphere's ability to adjust rapidly to local conditions of solar heat input; the lack of oceans, which on Earth have a large resistance to temperature changes; the great range in altitude of the surface (see below *Character of the surface*); the strong internal heating of the atmosphere because of suspended dust; and the seasonal deposition and release of a large part of the Martian atmosphere at the poles.

The only direct measurements of wind speeds were made by the Viking and Pathfinder landers. Near-surface winds at these sites were usually regular in behaviour and generally light. Average speeds were typically less than 2 metres per second (4.5 miles per hour), although gusts up to 40 metres per second were recorded. Other observations, including streaks of windblown dust and patterns in dune fields and in the many varieties of clouds, have provided additional clues about surface winds.

Global circulation models, which incorporate all the factors understood to influence the behaviour of the atmosphere, predict a strong dependence of winds on the Martian seasons because of the large horizontal temperature gradients associated with the edge of the polar caps in the fall and winter. Strong jet streams with eastward velocities above 100 metres per second form at winter high latitudes. Circulation is less dramatic in spring and fall, when light winds predominate everywhere. On Mars, unlike on Earth, there is also a relatively strong north-south circulation that transports the atmosphere to and from the winter and summer poles. The general circulation pattern is occasionally unstable and exhibits large-scale wave motions and instabilities: a regular series of rotating high- and low-pressure systems was clearly seen in the pressure and wind records at the Viking lander sites.

Local and global dust storms

Turbulence is an important factor in raising and maintaining the large quantity of dust found in the Martian atmosphere. Dust storms tend to begin at preferred locations in the southern hemisphere during the southern spring and summer. Activity is at first local and vigorous (for reasons yet to be understood), and large amounts of dust are thrown high into the atmosphere. If the amount of dust reaches a critical quantity, the storm rapidly intensifies, and dust is carried by high winds to all parts of the planet. In a few days the storm obscures the entire surface, and visibility is reduced to less than 5 percent of normal. The intensification process is evidently short-lived, as atmospheric clarity begins to return almost immediately, becoming normal typically in a few weeks.

THE POLAR CAPS

Seasonal changes. The seasonal behaviour of the Martian polar caps is well documented. Early each southern spring, the southern carbon dioxide cap starts to recede from its maximum extent of 50° S. As spring progresses, the cap shrinks by as much as 1° of latitude every five days. The edge of the cap becomes ragged, controlled by local topography (*e.g.*, craters), and eventually it breaks into well-defined fragments. From year to year, the location of the fragments remains about the same, although there are small variations in detail. Over the course of one-third of the Martian year, the cap recedes to its smallest extent; it is then not usually visible from Earth, but spacecraft observations have confirmed the presence of a small remnant throughout the summer.

Regrowth of the southern cap begins, after a brief period of atmospheric clarity, with the rapid formation of obscuring clouds, called the polar hood. On occasion the hood is sufficiently transparent to red light to let spacecraft view the formation of the cap. These limited opportunities, however, have not allowed its rate of advance toward the equator to be accurately known. The hood is far more extensive than the cap itself and can reach to within 35° of the equator. It probably contains particles of frozen water and carbon dioxide.

Differences in the behaviour of the northern carbon dioxide cap—including its smaller maximum size in winter and complete disappearance in summer—are due partly to differences in the lengths of seasons, in the distances from the Sun, and in elevation between the two poles.

Early debate over nature of polar caps

Composition of the caps. The composition of the seasonal polar caps was the subject of debate for nearly 200 years. One early hypothesis—that the caps were made of water ice—can be traced to Herschel, who imagined them to be just like those on Earth. In 1898 an Irish scientist, George J. Stoney, questioned this theory and suggested that the caps might consist of frozen carbon dioxide, but evidence to support the idea was not available until Kuiper's 1947 detection of carbon dioxide in the atmosphere.

In 1966 the American scientists Robert Leighton and Bruce Murray published the results of a numerical model of the thermal environment on Mars that raised considerable doubt about the water-ice hypothesis. Their calcula-

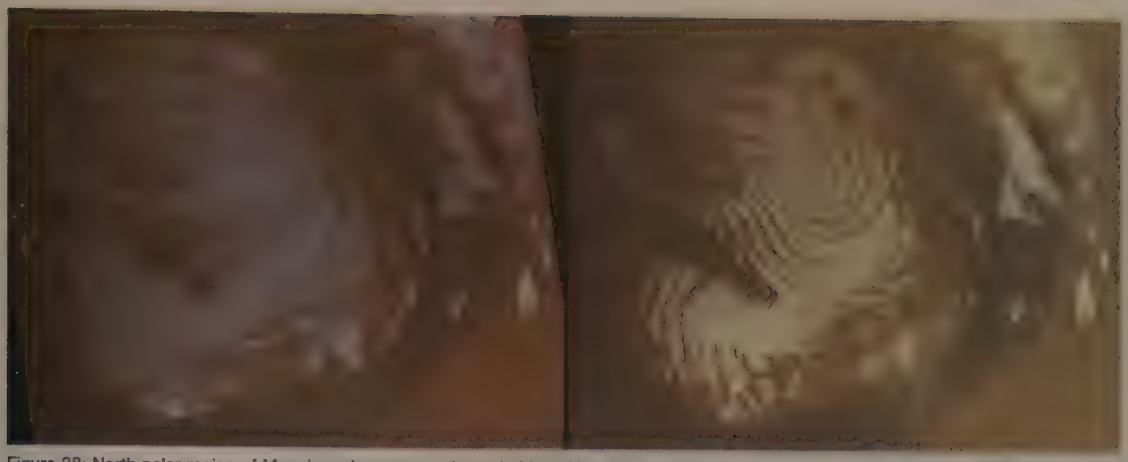


Figure 38: North polar region of Mars in early summer, shown in Mars Global Surveyor images made a Martian year apart. Visible are the bright remnant water-ice cap, distinctive spiral-patterned escarpments and valleys, and surrounding dune field (dark band) that occupies the northern part of the Vastitas Borealis plain. Comparison of the images reveals distinct variations in the details of the frost cover, suggesting substantial changes in the region's heat budget from year to year.
NASA/JPL/Malin Space Science Systems

tions indicated that, under Martian conditions, atmospheric carbon dioxide would freeze at the poles, and the growth and shrinkage of their model carbon dioxide caps mimicked the observed behaviour of the actual caps. The model predicted that the seasonal caps were relatively thin, only a few metres deep near the poles and thinning toward the equator. Although based on simplifications of the actual conditions on Mars, their results were later confirmed by measurements taken by the twin Mariner 6 and 7 spacecraft when they flew by Mars in 1969.

The composition of the summer remnant caps, particularly the southern one, remains somewhat less certain despite considerable data on their thermal and radiative properties. Measurements by the Viking orbiters showed that the ice of the northern remnant is unquestionably frozen water. Added to this evidence is the large increase in the amount of water vapour detected in the atmosphere over the summer cap. The northern remnant cap, in fact, represents the largest known reservoir of available water on the planet. Observations of the southern remnant cap have given more ambiguous results, although the cap is probably mostly carbon dioxide. Almost no water vapour has been observed in the atmosphere above it.

Polar terrain. The terrain in the polar regions is among the most distinctive on Mars. Beneath the seasonal and remnant caps at both poles, stacks of layered deposits up to three kilometres thick extend out to about the 80° latitude circle. The layers are exposed in escarpments and valleys that have a distinctive spiral pattern. Their composition is not known, but they likely include substantial amounts of water ice and dust. The layering is thought to result from climate-caused variations in the deposition rates of dust and water ice, which are traceable to changes in the planet's orbit and rotation. The layered terrains in the north lack impact craters, suggesting that they are very young. In contrast, cratering of the layered terrains in the south indicate an age of roughly 100 million years.

The great northern dune band

The north polar region also contains the largest area of sand dunes on Mars. The dunes, which occupy the northern part of the plain known as Vastitas Borealis, form a band that almost completely encircles the north polar remnant cap. Interlayering of sand and seasonal carbon dioxide snow can be seen in some locations, indicating that the dunes are active on at least a seasonal timescale.

CHARACTER OF THE SURFACE

The character of the Martian terrain has been well established from spacecraft photography and altimetry. The entire planet was photographed by the Viking orbiters at a resolution of roughly 250 metres and selected areas at resolutions down to 10 metres. Subsequently, the camera on the Mars Global Surveyor spacecraft photographed selected areas with resolutions of 1.4 metres, but it covered only a small fraction of the planet. The topography of the Martian surface was determined very accurately with the laser altimeter aboard Mars Global Surveyor, which mapped elevations with a vertical resolution of a few metres.

Lowest and highest points on Mars

Despite its small size, Mars has more relief than Earth. The lowest point on the planet, within the Hellas impact basin, is 8 kilometres below the reference level. The highest point, at the summit of the volcano Olympus Mons, is 21 kilometres above the reference level. The elevation range is thus 29 kilometres, compared with about 20 kilometres on Earth—*i.e.*, from the bottom of the Mariana Trench to the top of Mount Everest. Because Mars has no oceans, a reference level for elevations had to be defined in terms other than sea level. In the early 1970s the elevation at which the atmospheric pressure is 6.1 millibars (about 0.006 of the sea-level pressure on Earth) was set as the reference. When Mars Global Surveyor acquired more accurate elevation data, a better reference was needed, and the planet's mean radius of 3,389.51 kilometres was chosen.

One of the most striking aspects of the Martian surface is the contrast between the southern and northern hemispheres. Most of the southern hemisphere is high-standing and heavily cratered, resembling the battered highlands of the Moon. Most of the northern hemisphere is low-lying and sparsely cratered. The difference in mean elevation between the two hemispheres is roughly six kilometres. The topographic boundary between the hemispheres is not parallel to the equator but roughly follows a great circle inclined to it by about 30°. In some places the boundary is broad and irregular; in other places there are steep cliffs. Some of the most intensely eroded areas on Mars occur along the boundary. Landforms there include outflow channels, areas of collapse called chaotic terrain, and an enigmatic mix of valleys and ridges known as fretted terrain. Straddling the two hemispheres on one side of the planet is the Tharsis rise, a vast volcanic dome standing 8 kilometres above Mars's mean radius, 12 kilometres above the northern plains, and more than 2 kilometres above the surrounding cratered southern highlands. On or near the Tharsis rise are the planet's largest volcanoes (see below *Tharsis and Elysium*).

Southern cratered highlands. The number of very large craters in the southern highlands implies a substantial age for the surface. Planetary scientists have established from lunar samples returned by Apollo missions that the rate of large asteroid impacts on the Moon declined rapidly between 3.8 billion and 3.5 billion years ago. Surfaces that formed before this time are heavily cratered; those that formed after are less so. Mars very likely had a similar cratering history. Thus, the southern highlands probably formed more than 3.5 billion years ago.

The southern terrain possesses several distinctive types of craters—huge impact basins; large, flat-bottomed craters;

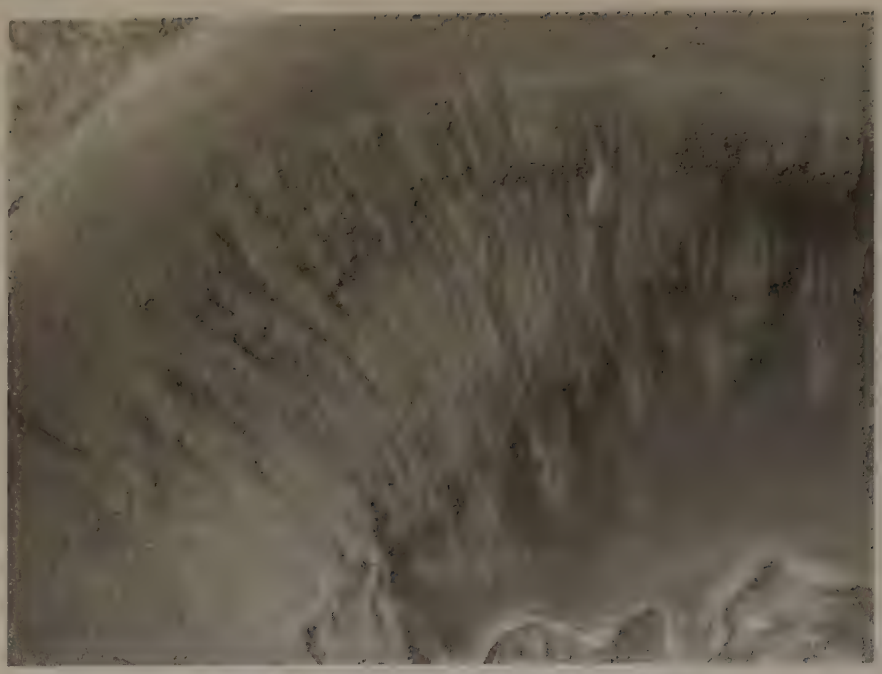


Figure 39: Fresh-appearing gullies on the north wall of a small crater within Mars's Newton Crater (near 41° S, 160° W), in a high-resolution Mars Global Surveyor image. These erosion features and similar ones found elsewhere on the planet suggest geologically recent water flow, but the interpretation is controversial.
NASA/JPL/Malin Space Science Systems

smaller, fresh-looking bowl-shaped craters like those on the Moon; and rampart and pedestal craters. Hellas is the largest impact basin on Mars. According to Mars Global Surveyor altimetry data, it is about 7,000 kilometres across and 8 kilometres deep—much larger than previously thought. Most of the craters measuring tens to hundreds of kilometres across are highly eroded. Because larger craters tend to be older than smaller craters, erosion rates on early Mars appear to have been much higher than subsequently. It is one piece of evidence that the climate on early Mars was very different from what it was for most of the planet's subsequent history.

Unique crater types

Rampart and pedestal craters may be unique to Mars. A rampart crater is so named because the lobes of ejecta—the material thrown out from the crater and extending around it—are bordered with a low ridge, or rampart. The ejecta thus apparently flowed across the ground, which may indicate that it had a mudlike consistency. Some scientists have conjectured that the mud formed from a mixture of impact debris and water that was present under the surface. Around a pedestal crater, the ejected material forms a steep-sided platform, or pedestal, with the crater situated inside its border. The pedestal appears to have developed when wind carved away the surface layer of the surrounding region while leaving intact that portion protected by the overlying ejecta.

High-resolution Viking images revealed an additional characteristic of the ancient southern terrain—the pervasive presence of networks of small valleys that resemble terrestrial drainage systems created by flowing water. Examples include Nirgal Vallis, located in the southern hemisphere north of the Argyre impact basin, and Nanedi Vallis, located just north of the equator near the east end of Valles Marineris. Scientists have proposed two alternative mechanisms for their formation, either the runoff of rainfall on the surface or erosion by groundwater that seeped onto the surface. In either case, warm climatic conditions were almost certainly required for their formation.

Fresh-appearing gullies

A major surprise of the Mars Global Surveyor mission was observations of small, fresh-appearing gullies on steep slopes at high latitudes. These features strongly resemble water-worn gullies in desert regions of the Earth, but their origin remains controversial. Although the discoverers initially proposed water erosion as the cause, this was challenged by other researchers.

Northern plains. The northern plains have remarkably little relief. They encompass all of the terrain within 30° of the pole except for the layered terrains immediately around the pole. Three broad lobes extend to lower latitudes. These include Chryse Planitia and Acidalia Planitia (centred on 30° W longitude), Amazonis Planitia (160° W), and Utopia Planitia (250° W). The only significant relief in this huge area is a large ancient impact basin, informally called the Utopia basin (40° N, 250° W).

Several different types of terrain have been recognized within the plains. In knobby terrain, numerous small hills are separated by smooth plains. The hills appear to be remnants of an ancient cratered surface now almost completely buried by younger material that forms the plains. Various plains have a polygonal fracture pattern that resembles landforms found in permafrost regions on Earth. Others have a peculiar thumbprint-like texture.

The origin of the low-lying northern plains remains controversial. Some scientists have proposed that they were formerly occupied by ocean-sized bodies of water that were fed by large floods (see below *Outflow channels*). Others have seen little evidence of global-sized bodies of water and have noted the difficulty of explaining the disappearance of such large water volumes.

Floods on early Mars

Outflow channels. Large flood channels, termed outflow channels, are observed incised into the Martian surface in several areas. The channels are generally tens of kilometres across and hundreds of kilometres long. Most emerge full-size from rubble-filled depressions and continue downslope into the northern plains or the Hellas basin in the south. Many of the largest drain from the south and west into Chryse Planitia. These are true channels in that they were once completely filled with flowing water, as opposed to most river valleys, which have never been close to full but contain a much smaller river channel. The peak discharges of the floods that cut the channels are estimated to have been 100–1,000 times the peak discharge of the Mississippi River. Some of the floods appear to have formed by catastrophic release of water from lakes. Others formed by explosive eruption of groundwater.

Valles Marineris. Close to the equator, centred on 70° W longitude, are several interconnected canyons collectively called Valles Marineris. Individual canyons are roughly 200 kilometres across. At the centre of the system, several canyons merge to form a depression 600 kilometres

across and as much as 9 kilometres deep—about five times the depth of the Grand Canyon. The entire system is more than 4,000 kilometres in length, or about 20 percent of Mars's circumference. At several places within the canyons are thick sedimentary sequences, which suggest that lakes may have formerly occupied the canyons. Some of the lakes may have drained catastrophically to the east to form large outflow channels that start at the canyons' eastern end. How the canyons formed is not known, but faulting probably played a major role.

Tharsis and Elysium. The canyons of Valles Marineris terminate to the west near the crest of the Tharsis rise, a vast bulge on the Martian surface more than 8,000 kilometres across and 8 kilometres high at its centre. Near the top of the rise are three of the planet's largest volcanoes—Ascraeus Mons, Arsia Mons, and Pavonis Mons—which tower 18, 17, and 14 kilometres, respectively, above the mean radius. Just off the rise to the northwest is the planet's tallest volcano, Olympus Mons, and at the north end is yet another large volcano, Alba Patera (approaching 7 kilometres high). What formed Tharsis is not known; it may have resulted from a combination of uplift and the accumulation of huge volumes of volcanic deposits.

Largest volcanoes on Mars

Another rise is located in the northern region of Elysium at about 215° W longitude. Much smaller than Tharsis, being only 2,000 kilometres across and 6 kilometres high, the Elysium rise is also the site of several volcanoes.

THE INTERIOR

The interior of Mars is poorly known. Planetary scientists have yet to conduct a successful seismic experiment via spacecraft that would provide direct information on internal structure and so must rely on indirect inferences. The moment of inertia of Mars indicates that it has a central core with a radius of 1,300–2,000 kilometres. Isotopic data from meteorites determined to have come from Mars (see below *Meteorites from Mars*) demonstrate unequivocally that the planet differentiated—separated into a metal-rich core and rocky mantle—at the end of the planetary accretion period 4.5 billion years ago. The planet has no detectable magnetic field that would indicate convection (heat-induced flow) in the core today. Large regions of magnetized rock have been detected in the oldest terrains, however, which suggests that very early Mars did have a magnetic field but that it disappeared as the planet cooled and the core solidified.

Evidence for ancient magnetic field

Mars is almost certainly volcanically active today, although at a very low level. Some Martian meteorites, which are all volcanic rocks, show ages as young as a few hundred million years, and some volcanic surfaces on the planet are so sparsely cratered that they must be only tens of millions of years old. Thus Mars was volcanically active in the geologically recent past, which implies that its mantle is warm and undergoing melting locally.

Mars's gravitational field is very different from Earth's. On Earth, excesses and deficits of mass in the surface crust, owing to the presence of large mountains and ocean deeps, respectively, tend to be offset by compensating masses at depth (isostatic compensation). Thus, the pull of gravity on Earth is the same on high mountains as it is over the ocean. This is also true for Mars's oldest terrains, such as the Hellas basin, the southern highlands, and the northern plains. The younger terrains such as the Tharsis and Elysium domes, however, are only partly compensated. Associated with both of these regions are gravity highs—that is, places where the measured gravity is significantly higher than elsewhere because of the large mass of the domes.

Because the gravity over the southern highlands is roughly the same as that over the low-lying northern plains, the southern highlands must be underlain by a thicker crust of material that is less dense than the mantle below it. Estimates of the thickness of the Martian crust range from only 3 kilometres under the Isidis impact basin, which is just north of the equator and east of Syrtis Major, to more than 90 kilometres at the south end of the Tharsis rise.

METEORITES FROM MARS

By 2002, scientists had identified more than 20 meteorites that have come from Mars. Suspicions about their origin were first raised when some meteorites that appeared to be volcanic rocks were found to have ages of about 1.3 billion years instead of the 4.5 billion years of all other meteorites. These rocks had to have come from a body that was geologically active in the comparatively recent past, and Mars was the most likely candidate. The rocks also have similar ratios of oxygen isotopes, which are distinctively different from those for Earth rocks, lunar rocks, and other meteorites. A Martian origin was finally proved when it was found that several of them contained trapped gases having a composition identical to that of the Martian atmosphere as measured by the Viking landers. The rocks are thought to have been ejected from the Martian surface by large impacts. They then went into solar orbit for several million years before falling on Earth. Claims in the mid-1990s of finding evidence for past microscopic life in one of the meteorites have been viewed skeptically by the general science community (see below *The question of life on Mars*).

THE SATELLITES

Little was learned about the two moons of Mars, Phobos and Deimos, after their discovery in 1877 until orbiting spacecraft observed them a century later. Viking 1 flew to within 100 kilometres of Phobos, and Viking 2 to within 30 kilometres of Deimos.

Phobos revolves around Mars once every 7 hours 39 minutes. It moves in an exceptionally close orbit, at a mean distance from the surface of about 6,000 kilometres—less than twice the planet's radius. Gravitational (tidal) forces slow the motion of Phobos and may ultimately cause the satellite to collide with Mars, possibly in less than 100 million years. Deimos suffers the opposite fate. It moves in a more distant orbit, and tidal forces are causing it to recede from the planet. Phobos and Deimos are not visible from all locations on the planet because of their small size, proximity to Mars, and near-equatorial orbits.

Close orbit of Phobos

Both moons are irregular chunks of rock, roughly ellipsoidal in shape. Phobos is the larger (see Table 11). Phobos's rugged surface is covered with impact craters. The largest crater, Stickney, is about half as wide as the moon itself. Its surface also exhibits a widespread system of linear fractures, or grooves, many of which are geometrically related to Stickney. In contrast, the surface of Deimos appears smooth, as its many craters are almost completely buried by fine debris, and it shows no fracture system. The albedo, or reflectivity, of both moons is very low, similar to that of the most primitive types of meteorites. One theory of the origin of the moons is that they are asteroids that were captured when Mars was forming.

Table 11: Satellites of Mars

property	Deimos	Phobos
Orbital radius (mean distance from centre of planet)	23,459 km	9,378 km
Orbital period (sidereal period)	1.26244 Earth days	0.31891 Earth days
Mean orbital velocity	1.4 km/s	2.1 km/s
Rotation period	sync.*	sync.*
Dimensions	$15 \times 12.2 \times 10.4$ km	$26.6 \times 22.2 \times 18.6$ km
Area	525 km^2	1,625 km^2
Mass	1.8×10^{15} kg	1.08×10^{16} kg
Mean density	1.8 g/cm^3	1.9 g/cm^3
Mean escape velocity	6 m/s	10 m/s
Albedo	0.07	0.06

*Sync. = synchronous rotation; the rotation and orbital periods are the same.

SPACECRAFT EXPLORATION

Between 1960 and 1980, the exploration of Mars was a major objective of both the U.S. and Soviet space programs. U.S. spacecraft successfully flew by Mars (Mariners 4, 6, and 7), orbited the planet (Mariner 9 and Vikings 1 and 2), and placed lander modules on its surface (Vikings 1 and 2). Three Soviet probes (Mars 2, 3, and 5) also investigated Mars, two of them reaching its surface. Mars 3 was the first spacecraft to soft-land an instrumented capsule on the planet, on Dec. 2, 1971.

The central theme of the Viking missions was the search

Figure 40: Phobos (left) and Deimos (right), photographed by Vikings 1 and 2, respectively, in 1977. The smooth texture of the surface of Deimos is contrasted with the grooved, pitted, and cratered surface of Phobos. The large cavity on Phobos is Stickney crater.
NASA/Malin Space Science Systems

for extraterrestrial life. No unequivocal evidence of biological activity was found (see below *The question of life on Mars*), but various instruments on the two orbiters and two landers returned detailed information concerning Martian geology, meteorology, and the physics and chemistry of the upper atmosphere. Vikings 1 and 2 were placed into orbit in June and August 1976, respectively. Lander modules descended to the surface from the orbiters after suitable sites were found. Viking 1 landed in the region of Chryse Planitia (22° N, 48° W) on July 20, 1976, and Viking 2 landed 6,700 kilometres away in Utopia Planitia (48° N, 226° W) on Sept. 3, 1976.

Amid failures of several U.S. missions to Mars in the 1990s, Mars Pathfinder successfully set down in Chryse Planitia (19° N, 33° W) on July 4, 1997, and deployed a robotic wheeled surface rover called Sojourner. This was followed by Mars Global Surveyor, which reached Mars in September 1997 and systematically mapped various properties of the planet from orbit for several years beginning in March 1999. These included Mars's gravity and magnetic fields, surface topography, and surface mineralogy. The spacecraft also carried cameras for making both wide-angle and detailed images of the surface at resolutions down to 1.5 metres. In October 2001 Mars Odyssey entered Mars orbit and started mapping such properties as surface chemical composition, distribution of near-surface ice, and the physical properties of near-surface materials.

A wave of spacecraft converged on Mars in late 2003 and early 2004 with mixed outcomes. Nozomi, launched by Japan in 1998 on a leisurely trajectory, was first to reach the vicinity of the planet, but malfunctions prevented it from entering Mars orbit. In mid-2003 the European Space Agency's Mars Express was launched on a half-year journey to the Red Planet. Carrying instruments to study the atmosphere, surface, and subsurface, it entered Mars orbit on December 25; however, its lander, named Beagle 2,

Mars Express

NASA/JPL/Cornell University

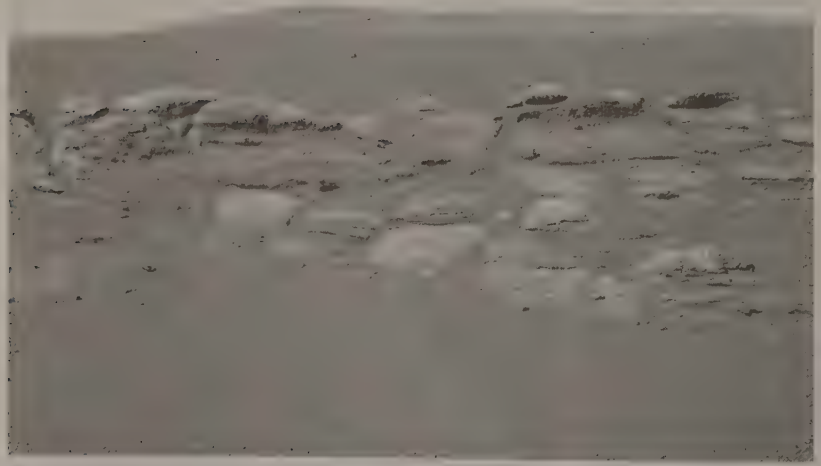


Figure 41: A portion of rock outcropping within a small crater in the Meridiani Planum region of Mars, shown in an image made by the robotic rover Opportunity in January 2004. The outcropping varies in height from 30 to 45 centimetres (12 to 18 inches) along the crater. As interpreted by mission scientists, the rock layers apparently were laid down as deposits at the bottom of a body of flowing salt water, probably on the shoreline of an ancient sea.

which was to examine the rocks and soil for signs of past or present life, failed to establish radio contact after presumably descending to the Martian surface the same day. Within weeks of its arrival, the Mars Express orbiter detected vast fields of water ice as well as carbon dioxide ice at the south pole and confirmed that the southern summer remnant cap contains permanently frozen water.

Spirit and Opportunity rovers

Also launched in mid-2003 was the U.S. Mars Exploration Rover Mission, which comprised twin robotic landers, Spirit and Opportunity. Spirit touched down in Gusev crater (15° S, 175° E) on Jan. 3, 2004. Three weeks later, on January 24, Opportunity landed in Meridiani Planum (2° S, 6° W), on the opposite side of the planet. The wheeled rovers, each carrying cameras and a suite of instruments, analyzed the rocks, soil, and dust around their landing sites, which had been chosen because they appeared to have been affected by water in Mars's past. Both rovers found evidence of past water; perhaps the most dramatic was the discovery by Opportunity of rocks that appeared to have been laid down at the shoreline of an ancient body of salty water.

THE MAPPING OF MARS

The first known map of Mars was produced in 1830 by Wilhelm Beer and Johann Heinrich von Mädler of Germany. The Italian astronomer Giovanni Schiaparelli prepared the first modern astronomical map of Mars in 1877, which contained the basis of the system of nomenclature still in use today. The names on his map are in Latin and are formulated mostly in terms of the ancient geography of the Mediterranean area. This map also showed, for the first time, indications of an interconnecting system of straight lines on the bright areas that he described as *canali* (Italian: "channels"). Schiaparelli is usually credited with their first description, but his fellow countryman Pietro Angelo Secchi developed the idea of *canali* about a decade earlier. In the late 19th century the American astronomer Percival Lowell established an observatory in Flagstaff, Arizona, to observe Mars and produced ever more elaborate maps of the Martian canals until his death in 1916.

Observations made by Mars-orbiting spacecraft have led to many new maps presenting topography, geologic provinces, temperatures, mineral distribution, and a variety of other data. The prime meridian on Mars, the equivalent of the Greenwich meridian on Earth, is defined by a small crater named Airy-0 located in the crater Airy.

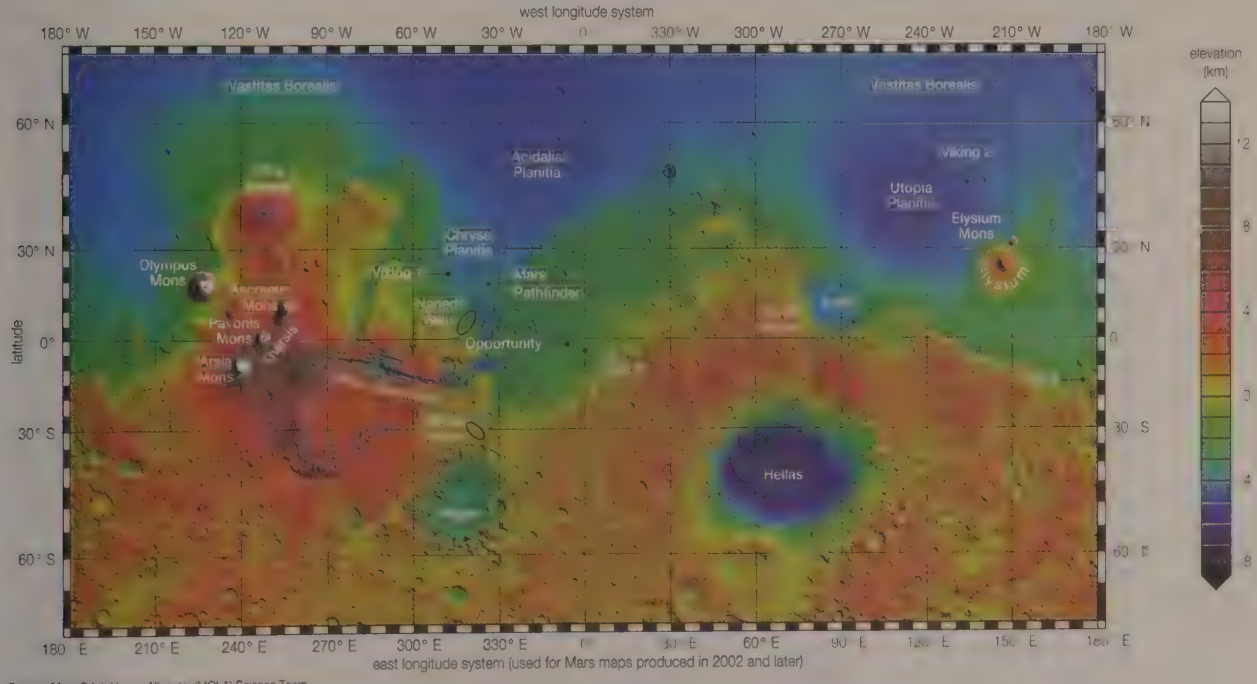
THE QUESTION OF LIFE ON MARS

The possible presence of life on Mars has been an essential element of general discussions of the planet since Schiaparelli first included *canali* on his maps. Lowell was particularly responsible for popularizing the notion that these markings were the result of biological activity, intelligent or otherwise. Nevertheless, as discussed above, it has been clear since the 1960s that such features do not exist.

Viking biological experiments

The biological experiments aboard the two Viking landers addressed three issues: (1) the nature of organic material, if any, on the Martian surface, (2) the possible presence of objects on the surface whose appearance or motion would suggest living or fossilized organisms, and (3) the possible presence in Martian soil of agents that, under prescribed conditions, could indicate metabolic processes. The results related to the first issue were unambiguous—a direct, extremely sensitive chemical analysis of samples at both lander sites showed no trace of complex organic materials. Addressing the second issue, the cameras on the landers found no evidence of biological agents or activity.

Three separate instruments addressed the last issue. One was designed to look for signs of photosynthesis or chemosynthesis—that is, signs of biological activity supported by solar or chemical energy, respectively—in samples of Martian soil. This experiment produced some positive indications, but the experimenters believed that they could be best explained by nonbiological processes. A second experiment measured gases released from a soil sample as it was exposed to a humid atmosphere or treated with a solution of organic nutrients. This experiment also produced a positive result in that the samples liberated oxygen in response to the nutrient. This reaction, however, also occurred even after samples had been baked on site at 145° C (293° F) for three hours, leading experimenters to conclude that the source of oxygen was nonbiological. The final experiment looked for the release of radioactive gas when a soil sample was exposed to a solution of radioactive organic nutrient. A positive result was again obtained, and in this case a baked control sample remained inert. Nevertheless, given the results of the other Viking experiments, most investigators believe that the re-



Source: Mars Orbital Laser Altimeter (MOLA) Science Team

Figure 42: Global topographic map of Mars produced from high-resolution laser altimetry data collected by Mars Global Surveyor through October 2000. This Mercator projection extends to latitudes 70° north and south. Topographic relief is colour-coded according to the key at right; selected major features of the planet are labeled.

Encyclopædia Britannica, Inc

sults of the labeled release experiment also can be explained nonbiologically.

The Martian surface is almost certainly devoid of life at the present time. It is exposed to ultraviolet radiation from the Sun without any reduction in intensity by the atmosphere; organic compounds in the soil are destroyed, probably by a combination of oxidation and photochemical processes; and average temperatures are so cold and the water content of the atmosphere so low that liquid water, universally accepted as essential for life, is unlikely to be available. These considerations have encouraged scientists to shift their search for life on Mars to the search for past life. As was indicated above, different lines of evidence suggest that conditions on early Mars were much more hospitable to life than subsequently. If life did gain a foothold, it now may survive only in protected niches well below the hostile surface. The current strategy, therefore, is to focus on evidence of past life and then look for present life deep below the surface, where liquid water might be available.

Purported fossils in a Martian meteorite

This strategy was underscored by the findings from the Martian meteorite ALH84001, which was collected from an Antarctic ice field in 1984. It is the oldest known rock from Mars, having formed about 4.5 billion years ago when the planet itself was accreting. It contains organic materials, evidence of mineral deposition from liquid water, and a number of structures similar to primitive terrestrial fossils. Although most scientists conclude that these findings do not provide good evidence for past life, they have stimulated closer examination of the Martian meteorites and have emphasized the need for obtaining samples from Mars, particularly of ancient rocks that were being formed or already present when conditions on Mars were more Earthlike. (M.J.S.B./M.C.Ma./M.H.Ca.)

Figure 43: Jupiter, photographed by Voyager 1 on Feb. 1, 1979, at a range of 32.7 million kilometres.
By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

Jupiter

Jupiter, designated ♃ in astronomy, is the most massive of the planets of the solar system and fifth in distance from the Sun. It is one of the brightest objects in the night sky; only the Moon, Venus, and sometimes Mars are more brilliant. When ancient astronomers named the planet for the Roman ruler of the gods and heavens (also known as Jove), they had no idea of the planet's true dimensions, but the name is appropriate, for Jupiter is larger than all the other planets combined. It takes nearly 12 years to orbit the Sun, and it rotates once about every 10 hours, more than twice as fast as Earth; its colourful cloud bands can be seen even with a small telescope. It has a narrow system of rings and at least 39 moons, one larger than the planet Mercury and three larger than Earth's Moon. Some astronomers speculate that Jupiter's moon Europa may be hiding an ocean of warm water—and possibly even some kind of life—beneath an icy crust.

Jupiter has an internal heat source—*i.e.*, it emits more energy than it receives from the Sun. The pressure in its deep interior is so high that the hydrogen there exists in a metallic state. This giant has the strongest magnetic field of any planet and a magnetosphere so large that, if it could be seen from Earth, its apparent diameter would exceed that of the Moon. Jupiter's system is also the source of intense bursts of radio noise, at some frequencies occasionally radiating more energy than the Sun. Despite all its superlatives, however, Jupiter is made almost entirely of only two elements, hydrogen and helium, and its mean density is not much more than the density of water.

Galileo orbiter and probe mission

Knowledge about the Jovian system grew dramatically after the mid-1970s as a result of explorations by three spacecraft missions—Pioneers 10 and 11 in 1973–74, Voyagers 1 and 2 in 1979, and the Galileo orbiter and probe, which arrived at Jupiter in December 1995. The Pioneer spacecraft served as scouts for the Voyagers, showing that the radiation environment was tolerable and mapping out the main characteristics of the planet. The greater number and increased sophistication of the Voyager instruments provided so much new information that it was still being analyzed when the Galileo mission began. The previous missions had all been flybys, but Galileo released a probe into Jupiter's atmosphere and then went into orbit about the planet for intensive investigations of the entire system over several years. Yet another view of the Jovian system was provided in 2000 by the flyby of the Cassini spacecraft on its way to Saturn. Observations of the impacts of the fragmented nucleus of Comet Shoemaker-Levy 9 with Jupiter's atmosphere in 1994 also yielded information about its composition and structure.

BASIC ASTRONOMICAL DATA

Jupiter has an equatorial diameter of about 143,000 kilometres and orbits the Sun at a mean distance of 778 million kilometres. Table 12 shows additional physical and orbital data. Of special interest are the planet's low mean density of 1.33 grams per cubic centimetre—in contrast with Earth's 5.52 grams per cubic centimetre—coupled with its large dimensions and mass and short rotation period. The low density and large mass indicate that Jupiter's composition and structure are quite unlike those of Earth and the other inner planets.

Three rotation periods, all within a few minutes of each other, have been established. The two periods called System I (9 hours 50 minutes 30 seconds) and System II (9 hours 55 minutes 41 seconds) refer to the speed of rotation

Table 12: Planetary Data for Jupiter

Mean distance from Sun	778,570,000 km (5.2 AU)
Eccentricity of orbit	0.049
Inclination of orbit to ecliptic	1.3°
Jovian year (sidereal period of revolution)	11.86 Earth years
Equatorial radius*	71,492 km
Polar radius*	66,854 km
Mass	18.99×10^{26} kg
Mean density	1.33 g/cm^3
Gravity*	2,312 cm/s^2
Escape velocity	59.5 km/s
Rotation periods	
System I (±10° from equator)	9 h 50 min 30 s
System II (higher latitudes)	9 h 55 min 41 s
System III (magnetic field)	9 h 55 min 29 s
Mean synodic period†	398.88 Earth days
Inclination of equator to orbit	3.13°
Dimensions of Great Red Spot	20,000 × 12,000 km
Mean opposition visual magnitude	−2.70
Magnetic field strength at equator	4.3 gauss
Number of known moons	39

*Calculated for the altitude at which 1 bar of atmospheric pressure is exerted. †Time required for the planet to return to the same position in the sky relative to the Sun as seen from Earth.

at the equator and at higher latitudes, respectively, as exhibited by features observed in the planet's visible cloud layers. Jupiter has no solid surface; the transition from the gaseous atmosphere to the fluid interior occurs gradually at great depths. Thus the variation in rotation period at different latitudes does not imply that the planet itself rotates with either of these mean velocities. In fact, the true rotation period of Jupiter is System III (9 hours 55 minutes 29 seconds), which is the period of rotation of Jupiter's magnetic field, first deduced from Earth-based observations at radio wavelengths and confirmed by direct spacecraft measurements. This period applies to the massive interior of the planet, where the magnetic field is generated.

THE OUTER LAYERS

The clouds and the Great Red Spot. Even a modest telescope can show much detail on Jupiter. The region of the planet's atmosphere that is visible from Earth contains several different types of clouds that are separated both vertically and horizontally. Changes in these cloud systems can occur over periods of hours, but an underlying pattern of latitudinal currents has maintained its stability for decades. It has become traditional to describe the appearance of the planet in terms of a standard nomenclature for its alternating dark bands, called belts, and bright bands, called zones. The underlying currents, however, seem to persist longer than this pattern.

Stability of underlying currents

The close-up views of Jupiter transmitted from the Voyager spacecraft revealed a variety of cloud forms, including many elliptical features reminiscent of storm systems on Earth. All these systems are in motion, appearing and disappearing on timescales that vary with their sizes and locations. Also observed to vary are the pastel shades of colours present in the cloud layers—from the tawny yellow that seems to characterize the main layer, through browns and blue-grays, to the well-known salmon-coloured Great Red Spot, Jupiter's largest and longest-lived feature. Chemical differences in cloud composition are presumed to be the cause of the variations in colour.

Unlike Earth, where local weather is often influenced by the varied nature of the planet's surface, Jupiter has no solid surface. In the absence of topographic features, the planet's large-scale circulation is dominated by latitudinal currents. The lack of a solid surface with physical boundaries and regions with different heat capacities makes the persistence of these currents and their associated cloud patterns all the more remarkable. The Great Red Spot, for example, moves in longitude with respect to all three of the planet's rotation systems, yet it does not move in latitude. The white ovals found just south of the Great Red Spot exhibit similar behaviour; white ovals of this size are found nowhere else on the planet. The dark brown clouds, evidently holes in the tawny cloud layer, are found almost exclusively near 18° N latitude. The blue-gray or purple areas, from which the strongest thermal emission is detected, occur only in the equatorial region of the planet.

Nature of the Great Red Spot. The true nature of Jupiter's unique Great Red Spot was still unknown at the start of the 21st century, despite extensive observations from the Voyager and Galileo spacecraft. On a planet whose cloud patterns have lifetimes often counted in days, the Great Red Spot has survived as long as detailed observations of Jupiter have been made—at least 300 years. Its present dimensions are about 20,000 by 12,000 kilometres, making it large enough to accommodate both Earth and Mars. These huge dimensions are probably responsible for the feature's longevity and possibly for its distinct colour.

Voyager observations revealed that the material within the Great Red Spot rotates counterclockwise once every seven days, corresponding to superhurricane-force winds of 400 kilometres per hour at the periphery. The Voyager images also recorded a large number of interactions between the spot and much smaller disturbances moving in the current at the same latitude. The interior of the spot is remarkably tranquil, with no clear evidence for the expected upwelling of material from lower depths. The Great Red Spot, therefore, appears to be a huge anticyclone, a vortex or eddy whose diameter is presumably accompanied by a great depth that allows the feature to reach well below and well above the main cloud layers.

Cloud composition. Jupiter's visible clouds are formed at different altitudes in the planet's atmosphere. Except for the top of the Great Red Spot, the white clouds are the highest, with cloud-top temperatures of about 120 kelvins (K; −240° F, −150° C). These clouds consist of frozen ammonia crystals and are analogous to the water-ice cirrus clouds on Earth. The tawny clouds that are widely distributed over the planet occur at lower levels. They appear to form at a temperature of about 200 K (−100° F, −70° C), which suggests that they consist of condensed ammonium

Ammonia cirrus clouds

By courtesy of the Jet Propulsion Laboratory, National Aeronautics and Space Administration

Figure 44: The Great Red Spot (top right) and the surrounding region, photographed by Voyager 1 on March 1, 1979. At centre right is one of the associated white ovals.

Table 13: Atmospheric Abundances for Jupiter

equilibrium species				nonequilibrium species			
gas	mixing ratio*	element measured (relative to hydrogen)	Jupiter/Sun ratio	gas	mixing ratio*	element measured (relative to hydrogen)	Jupiter/Sun ratio
hydrogen (H_2)	1.0			phosphine (PH_3)	6×10^{-7}	phosphorus	0.8
helium (He)	0.16	helium-4	0.81	germane (GeH_4)	7×10^{-10}	germanium	0.05
water (H_2O)	$>3 \times 10^{-4}$	oxygen	>0.82	arsine (AsH_3)	2×10^{-10}	arsenic	0.5
methane (CH_4)	2.1×10^{-3}	carbon	2.9 ± 0.5	carbon monoxide (CO)	1.6×10^{-9}		
ammonia (NH_3)	8.1×10^{-4}	nitrogen	3.6 ± 0.5	carbon dioxide (CO_2)	detected in stratosphere		
hydrogen sulfide (H_2S)	7.7×10^{-5}	sulfur	2.5 ± 0.2	ethane (C_2H_6)	$1–5 \times 10^{-6}$ (stratosphere)		
hydrogen deuteride (HD)	5.2×10^{-5}	deuterium	no deuterium on Sun	acetylene (C_2H_2)	$3–10 \times 10^{-8}$ (stratosphere)		
neon (Ne)	2.5×10^{-5}	neon-20	0.10 ± 0.01	ethylene (C_2H_4)	7×10^{-9} (north polar region)		
argon (Ar)	1.8×10^{-5}	argon-36	2.5 ± 0.5	benzene (C_6H_6)	2×10^{-9} (north polar region)		
krypton (Kr)	8.7×10^{-9}	krypton-84	2.7 ± 0.5				
xenon (Xe)	8.8×10^{-10}	xenon-132	2.6 ± 0.5				

*The mixing ratio is the number of molecules of a given atmospheric constituent in a unit volume divided by the number of hydrogen molecules in that same volume.

hydrosulfide and that their colour is caused by other ammonia-sulfur compounds. Sulfur compounds are invoked as the likely colouring agents because sulfur is relatively abundant in the cosmos and hydrogen sulfide is notably absent from Jupiter's atmosphere above the clouds.

Jupiter is composed primarily of hydrogen and helium. Under equilibrium conditions—allowing all the elements present to react with one another at an average temperature for the visible part of the Jovian atmosphere—the abundant elements are all expected to combine with hydrogen. Thus it was surmised that methane, ammonia, water, and hydrogen sulfide would be present. Except for hydrogen sulfide, all these compounds have been found by spectroscopic observations from Earth. The apparent absence of hydrogen sulfide can be understood if it combines with ammonia to produce the postulated ammonium hydrosulfide clouds. The absence of detectable hydrogen sulfide above the clouds suggests that the chemistry that forms coloured sulfur compounds (if indeed there are any) must be driven by local lightning discharges rather than by ultraviolet radiation from the Sun.

Ammonia-sulfur clouds

Sulfur compounds have also been proposed to explain the dark brown coloration of the ammonia clouds detected at still lower levels, where the temperature is 260 K (8° F, −13° C). These clouds are seen through what are apparently holes in the otherwise ubiquitous tawny clouds.

The colour of the Great Red Spot has been attributed variously to the presence of complex organic molecules, red phosphorus, or yet another sulfur compound. Dark regions occur near the heads of white plume clouds near the planet's equator, where temperatures as high as 300 K (80° F, 27° C) have been measured. Despite their blue-gray appearance, they have a reddish tint. They appear to be cloud-free regions—hence the ability to "see" into them to great depths and measure high temperatures—that exhibit a blue colour (from Rayleigh scattering of sunlight) overlain with a thin haze of reddish material. That these so-called hot spots occur only near the equator, the elliptical dark brown clouds only near 18° north latitude, and the most prominent red colour on the planet only in the Great Red Spot implies a cloud chemistry whose localization is puzzling in such a dynamically active atmosphere.

Water clouds

At still lower depths, astronomers expect to find water-ice clouds and water-droplet clouds, both consisting of dilute solutions of ammonium hydroxide. Nevertheless, when the probe from the Galileo spacecraft entered Jupiter's atmosphere on Dec. 7, 1995, it failed to find these water clouds, even though it survived to a pressure level of 24 bars—nearly 24 times sea-level pressure on Earth—where the temperature was more than 400 K (260° F, 130° C). In fact, the probe also did not sense the upper cloud layers of ammonia and ammonium hydrosulfide. Unfortunately for studies of Jovian cloud physics, the probe had entered the atmosphere over a hot spot, where clouds were absent.

The atmosphere. *Proportions of constituents.* Before the Galileo mission, astronomers had relied on studies of the planet's spectrum to provide information about the composition, temperature, and pressure of the atmosphere. The presence of methane and ammonia was deduced in this way in the 1930s, while hydrogen was detected for the first time in 1960. Subsequent studies led to a growing list of new constituents, including the discovery of the arsenic compound arsine in 1990. Table 13 includes a list of Jupiter's atmospheric constituents and their abundances as determined by Earth-based, spacecraft, and atmospheric probe observations as of 2000.

If the condition of chemical equilibrium held rigorously in Jupiter's atmosphere, one would not expect to find molecules such as carbon monoxide or phosphine in the abundances measured. Neither would one expect the traces of acetylene, ethane, and other hydrocarbons that have been detected in the stratosphere. Evidently, there are sources of energy other than the molecular kinetic energy corresponding to local temperatures. Solar ultraviolet radiation is responsible for the breakdown of methane, and subsequent reactions of its fragments produce acetylene and ethane. In the convective region of the atmosphere, lightning discharges (observed by the Voyager and Galileo spacecraft) contribute to these processes. Still deeper, at temperatures around 1,200 K (1,700° F, 930° C), carbon monoxide is made by a reaction between methane and water vapour.

Galileo's probe carried a mass spectrometer that detected the constituent atoms and molecules in the atmosphere. As the probe descended through the atmosphere on its parachute, the spectrometer also studied variations in abundance with altitude. This experiment finally detected the previously missing hydrogen sulfide, which was found to be present even lower in the atmosphere than anticipated. Evidently this cloud-forming gas, like ammonia and water vapour, was depleted in the upper part of the aforementioned hot spot. It was not possible to measure oxygen, because this element is bound up in water, and the probe did not descend into the hot spot deeply enough to reach the region where this condensable vapour is well mixed.

Differences between Jupiter and the Sun

The elemental abundances in Jupiter's atmosphere can be compared with the composition of the Sun (see Table 13). If, like the Sun, the planet had formed by simple condensation from the primordial solar nebula that is thought to have given birth to the solar system, their elemental abundances should be the same. A surprising result from the Galileo probe was that all the elements that it could measure in the Jovian atmosphere showed the same approximately three-fold enrichment of their values in the Sun, relative to hydrogen. This has important implications for the formation of the planet (see below *Theories of the origin of the Jovian system*). Spectroscopy from Earth reveals a large spread in the values of other elements not measured by the probe. The abundances of the gases from which these elemental abundances are derived depend on dynamical phenomena in Jupiter's atmosphere—principally chemical reactions and vertical mixing. The significance of the helium depletion is discussed below in *The interior*.

Another difference with solar values is indicated by the presence of deuterium on Jupiter. This heavy isotope of

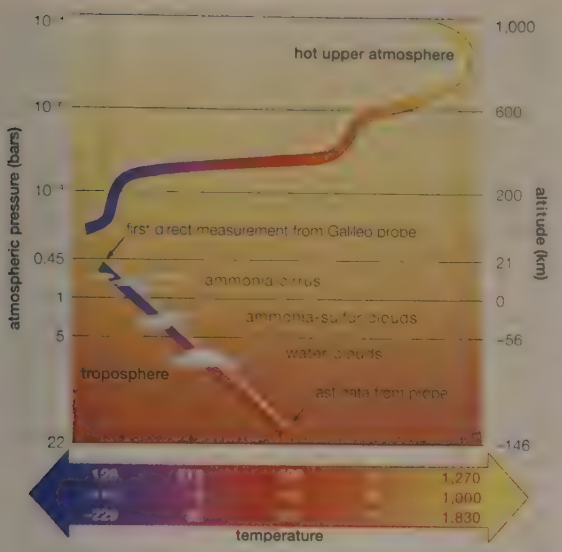


Figure 45: Profile of Jupiter's atmosphere as deduced from accelerometer data and direct measurements collected by the Galileo probe. The altitude reference level is set at one bar (sea-level pressure on Earth). Schematic clouds indicate the approximate positions of the expected cloud layers (see text).
EB Inc

hydrogen has disappeared from the Sun as a result of nuclear reactions in the solar interior, but no such reactions have occurred on Jupiter.

Temperature and pressure. The Galileo probe carried instruments to record both temperature and pressure during its atmospheric descent. This profile is illustrated in Figure 45, which includes the locations of the different cloud layers if they had occurred where expected.

It is notable in the profile that temperatures higher than the freezing point of water (273 K, 32° F, 0° C) were measured at pressures just a few times greater than sea-level pressure on Earth (about one bar). This is mainly a consequence of Jupiter's internal energy source, although some warming would occur just through the trapping of infrared radiation by the atmosphere. The temperature rise above the tropopause (upper part of Figure 45) is known as an inversion, because temperature normally decreases with height. The inversion is caused by the absorption of solar energy at these altitudes by gases and aerosol particles.

Other likely atmospheric constituents. The list of atmospheric abundances in Table 13 is certainly not complete. For example, astronomers expect monosilane to be present in the deep atmosphere, along with many other exotic species. Other nonequilibrium species should occur in the higher regions, accessible to future probes, as a result of chemical reactions driven by lightning or solar ultraviolet radiation, or at the poles (where, for example, benzene has been detected) by the precipitation of charged particles.

Relevance to origin-of-life research

The formation of complex organic molecules in Jupiter's atmosphere is of great interest in the study of the origin of life. The initial chemical processes that gave rise to living organisms on Earth may have occurred in transient microenvironments that resembled the present chemical composition of Jupiter, without the enormous amounts of hydrogen and helium. Thus, Jupiter may represent a vast natural laboratory in which the first steps toward the origin of life are being pursued again and again. Determining how complex prelife chemical processes can become under such conditions is one of the most fascinating problems confronting any program of space exploration.

Radio emission. Jupiter was the first planet found (in 1955) to be a source of radiation at radio wavelengths. The radiation was recorded at a frequency of 22 megahertz (corresponding to a wavelength of 13.6 metres, or 1.36 decametres). It took the form of noise bursts with peak intensities sometimes great enough to make Jupiter the brightest source in the sky at this wavelength, except for the Sun during its most active phase. These bursts constituted the first evidence for a Jovian magnetic field. Later observations at shorter (decimetre) wavelengths revealed that Jupiter is also a source of steady radio emission. It has become customary to refer to these two types of emission in terms of their characteristic wavelengths—decametre radiation and decimetre radiation.

The nonthermal component of the continuous, decimetre radiation is interpreted as synchrotron emission—that is, radiation emitted by extremely high-speed electrons moving in the planet's magnetic field within a toroidal, or doughnut-shaped, region surrounding Jupiter—a phenomenon closely analogous to that of Earth's Van Allen belts. The maximum emission occurs at a distance of two planetary radii from the centre of the planet and has been detected at 178–5,000 megahertz. The intensity of the emission and its plane of polarization vary with the same period, which can be explained if the axis of the planet's magnetic field is inclined by about 10° to the rotational axis. The period of these variations is the System III rotation period.

The radio emission at the intermittent, decametre wavelengths has been studied from Earth in the accessible range of 3.5–39.5 megahertz. Free of Earth's ionosphere, which blocks lower frequencies from reaching the surface, the radio-wave experiment on the Voyager spacecraft detected emissions from Jupiter down to 60 kilohertz, corresponding to a wavelength of five kilometres. The strength of the signal and the frequency of noise storms show a marked time dependence that led to the early detection of three sources, or emitting regions.

Effect of Io's position on radio emissions

The decametre noise storms are greatly affected by the position of Jupiter's moon Io in its orbit. For one source, events are much more likely to occur when Io is 90° from the position in which Earth, Jupiter, and Io are in a straight line than otherwise. The noise sources appear to be regions that lie in the line of sight toward the visible disk of the planet. The most promising explanation of this effect relates the emission to a small region of space linked to Io by magnetic field lines (a flux tube) that move with Io. Electrons moving in spirals around the magnetic field lines could produce the observed radiation. Interactions between these electrons and the Jovian ionosphere indeed were observed by the Voyager and Galileo spacecraft. The "footprint" of Io's flux tube on Jupiter's upper atmosphere can even be observed from Earth as a glowing spot associated with Jupiter's polar auroras.

The Jovian magnetosphere. The nonthermal radio emissions described above are the natural result of trapped charged particles interacting with Jupiter's magnetic field and ionosphere. Interpretation of these observations led to a definition of the basic characteristics of the planet's magnetosphere that was shown to be remarkably accurate by the Pioneer and Voyager spacecraft.

The basic magnetic field of the planet is dipolar in nature, generated by a hydromagnetic dynamo that is driven by convection within the electrically conducting outer layers of Jupiter's interior. The magnetic field strength at the equator is 4.3 gauss, compared with 0.3 gauss at Earth's surface. The axis of the magnetic dipole is offset by a tenth of Jupiter's equatorial radius of 71,500 kilometres from the planet's rotational axis, to which it is indeed inclined by 10°. The orientation of the Jovian magnetic field is opposite to the present orientation of Earth's field.

The magnetic field dominates the region around Jupiter in the shape of an extended teardrop. The round side of the teardrop faces the Sun, where the Jovian field repels the solar wind, forming a bow shock at a distance of about 3 million kilometres from the planet. Opposite the Sun, an immense magnetotail stretches out to the orbit of Saturn, a distance of 650 million kilometres. These dimensions make Jupiter's magnetosphere the largest permanent structure in the solar system, dwarfing the Sun's diameter of 1.4 million kilometres. Within this region, the most striking activity is generated by the moon Io, whose influence on the decametric radiation is discussed in the section above. An electric current of approximately five million amperes flows in the magnetic flux tube linking Jupiter and Io. This satellite is also the source of a toroidal cloud of ions, or plasma, that surrounds its orbit.

Size of Jupiter's magnetosphere

The auroras. Just as charged particles trapped in the Van Allen belts cause auroras on Earth when they crash

into the uppermost atmosphere near the magnetic poles, so do they also on Jupiter. Cameras on the Voyager and Galileo spacecraft succeeded in imaging ultraviolet auroral arcs on the nightside of Jupiter, while Earth-based observations have recorded infrared emissions from the H_3^+ ion at both poles and imaged the associated polar auroras. Evidently, protons (hydrogen ions, H^+) from the magnetosphere spiral into the planet's ionosphere along magnetic field lines, forming the excited H_3^+ ion as they crash into the atmosphere dominated by molecular hydrogen. The resulting emission produces the auroras. The relation of the ultraviolet and infrared auroras, the detailed interaction of Io's flux tube with Jupiter's ionosphere, and the possibility that ions from Io's torus are impinging on the planet's atmosphere remain active topics of research.

THE INTERIOR

The atmosphere of Jupiter constitutes a very small fraction of the planet. Because nothing can be directly observed below this thin outer layer, only indirect conclusions can be drawn about the composition of Jupiter's interior. The observed quantities with which astronomers can work are the atmospheric temperature and pressure, mass, radius, shape, rate of rotation, heat balance, and perturbations of satellite orbits and spacecraft trajectories. From these can be calculated the ellipticity—or deviation from a perfect sphere—of the planet and its departure from an ellipsoid shape. These latter quantities may also be predicted using theoretical descriptions, or models, for the internal distribution of material. Such models can then be tested by their agreement with the observations.

The basic difficulty in constructing a model that will adequately describe the internal conditions for Jupiter is the absence of extensive laboratory data on the properties of hydrogen and helium at pressures and temperatures that would exist near the centre of this giant planet. The central temperature is estimated to be close to 25,000 K (44,500° F, 24,700° C), to be consistent with an internal source of heat that allows Jupiter to radiate about twice as much energy as it receives from the Sun. The central pressure is in the range of 50–100 million atmospheres (about 50–100 megabars). At such tremendous pressures hydrogen is expected to be in a metallic state.

Metallic hydrogen

Despite the problems posed in establishing the properties of matter in these extreme conditions, the precision of the models has improved steadily. Perhaps the most significant early conclusion from them was the realization that Jupiter cannot be composed entirely of hydrogen; if it were, it would have to be considerably larger than it is to account for its mass. On the other hand, hydrogen must predominate, constituting at least 70 percent of the planet by mass, regardless of form—gas, liquid, or solid. The Galileo probe measured a proportion for helium of 24 percent by mass in Jupiter's upper atmosphere, compared with the 28 percent predicted if the atmosphere had the same composition as the original solar nebula. Because the planet as a whole should have that original composition, astronomers have concluded that some helium that was dissolved in the liquid hydrogen in the planet's interior has precipitated out of solution and sunk toward the planet's centre, leaving the atmosphere depleted of this gas.

Current models agree on a transition from molecular to metallic hydrogen at about a fourth of the distance down toward Jupiter's centre. It should be stressed that this is not a transition between a liquid and a solid but rather between two fluids with different electrical properties. No solid surface exists in any of these models, although most (but not all) models incorporate a dense core with a radius of 0.03–0.1 that of Jupiter (0.33–1.1 the radius of Earth).

The source of internal heat has not been completely resolved. The favoured explanation invokes a combination of the gradual release of primordial heat left from the planet's formation and the liberation of thermal energy from the precipitation of droplets of helium in the planet's deep interior. The lower helium abundance in Jupiter's atmosphere relative to the Sun (Table 13) supports this latter deduction. The first process is simply the cooling phase of the original "collapse" that converted potential energy to thermal energy at the time the planet accumulated its complement of solar nebula gas (see below *Theories of the origin of the Jovian system*).

THE SATELLITES AND RING SYSTEM

The first objects in the solar system discovered by means of a telescope—by Galileo in 1610—were the four brightest moons of Jupiter, now called the Galilean satellites. The fifth known Jovian moon, Amalthea, was also discovered by visual observation—by E.E. Barnard in 1892. All the other known satellites were found photographically or in spacecraft images. Jupiter's system of thin rings was detected in Voyager images in 1979.

Satellite groups. Data for the known Jovian moons are summarized in Table 14. Roman numerals are assigned to the first 16 known moons in order of their discovery. The orbits of the inner eight moons have low eccentricities and low inclinations—*i.e.*, the orbits are all nearly circular and in the plane of the planet's equator. Such moons are called "regular." The orbits of the moons beyond Callisto have much higher inclinations and eccentricities, making them "irregular." The two innermost satellites, Metis and Adrastea, are intimately associated with Jupiter's ring system, as sources of the fine particles and as gravitationally controlling "shepherds." There are probably additional members of each of these groups.

"Regular" and "irregular" moons

The Galilean satellites. Galileo proposed that the four Jovian moons he discovered in 1610 be named the Medicean stars, in honour of his patron, Cosimo II de' Medici, but they soon came to be known as the Galilean satellites. In order of increasing distance from the planet, they are called Io, Europa, Ganymede, and Callisto, for fig-

Table 14: Satellites of Jupiter*

name	numerical designation	year of discovery	mean distance from Jupiter (km)	sidereal period (days)†	orbital inclination (degrees)‡	orbital eccentricity‡	radius or radial dimensions (km)‡	mass (kg)	mean density (g/cm³)‡
Metis	XVI	1979	128,000	0.295	(0)	0.0	(20)	1×10^{17}	
Adrastea	XV	1979	129,000	0.298	(0)	(0.0)	$(12.5 \times 10 \times 7.5)$	2×10^{16}	
Amalthea	V	1892	181,000	0.498	0.4	0.003	$131 \times 73 \times 67$	7.5×10^{18}	3.10
Thebe	XIV	1979	222,000	0.675	(0.8)	0.015	(55×45)	8×10^{17}	
Io	I	1610	422,000	1.769	0.04	0.004	$1,830 \times 1,819 \times 1,815$	8.932×10^{22}	3.55
Europa	II	1610	671,000	3.551	0.47	0.009	1,565	4.800×10^{22}	3.04
Ganymede	III	1610	1,070,000	7.155	0.21	0.002	2,634	1.482×10^{23}	1.93
Callisto	IV	1610	1,883,000	16.689	0.51	0.007	2,403	1.076×10^{23}	1.83
Leda	XIII	1974	11,094,000	238.72	26	0.148	5	6×10^{15}	
Himalia	VI	1904	11,480,000	250.57	28	0.158	85	9.5×10^{18}	
Lysithea	X	1938	11,720,000	259.22	29	0.107	12	8×10^{16}	
Elara	VII	1905	11,737,000	259.65	25	0.207	40	8×10^{17}	
Ananke	XII	1951	21,200,000	631 R	147	0.169	10	4×10^{16}	
Carme	XI	1938	22,600,000	692 R	164	0.207	15	1×10^{17}	
Pasiphae	VIII	1908	23,500,000	735 R	145	0.378	18	3×10^{17}	
Sinope	IX	1914	23,700,000	758 R	153	0.275	14	8×10^{16}	

*Between 1999 and 2001, 23 additional moons (including one seen in 1975 and then lost) were discovered photographically in Earth-based observations. All are irregular, with large orbital radii, eccentricities, and inclinations; 21 have retrograde orbits. Rough size estimates based on brightness place them between 2 and 8 km in diameter. Assigned numerical designations, they await official names. †R following the quantity indicates a retrograde orbit. ‡Quantities given in parentheses are poorly known. Unspecified quantities are unknown.

ures closely associated with Jupiter in Greek mythology. The names were assigned by the German astronomer Simon Marius, Galileo's contemporary and rival, who likely discovered the satellites independently.

Although approximate diameters and spectroscopic characteristics of the Galilean moons had been determined from Earth-based observations, it was the Voyager missions that indelibly established these four bodies as worlds in their own right. The Galileo mission provided a wealth of additional data. Before Voyager, it was known that Callisto and Ganymede were both as large as or larger than the planet Mercury; that they and Europa had surfaces covered with water ice; that Io's orbit was surrounded by a torus of atoms and ions that included sodium, potassium, and sulfur; and that the inner two Galilean moons had mean densities much greater than those of the outer two. This density gradient from Io to Callisto resembles that found in the solar system itself and seems to result from the same cause (see below *Theories of the origin of the Jovian system*). The density values suggest that Io and Europa have a rocky composition similar to that of the Moon, whereas roughly 50 percent of Ganymede and Callisto must be made of a much less dense substance, with water ice being the obvious candidate.

Lack of global heating and differentiation

Callisto. The icy surface of this satellite is so dominated by impact craters that there are no smooth plains like the dark maria observed on the Moon. Unmodified by any upwelling of material from internal activity, its appearance is consistent with the absence of a differentiated interior. Evidently no tidally induced global heating and consequent melting occurred on Callisto, unlike the other three Galilean moons. In addition to the predominant water ice, solid carbon dioxide is present on the surface, and a very tenuous carbon dioxide atmosphere is slowly escaping into space. Callisto has a weak magnetic field induced by Jupiter's field that may imply the existence of a layer of liquid water below its icy crust.

Ganymede. Unlike Callisto, Ganymede, an equally icy satellite, reveals distinct patches of dark and light terrain. This contrast is reminiscent of the Moon's surface, but the answer to which terrain came first—dark or light—is exactly reversed. In contrast to the Moon, the dark regions on Ganymede are the older areas, showing the heaviest concentration of craters. The light regions are younger, revealing a complex pattern of parallel and intersecting ridges and grooves in addition to unusually bright impact craters. This manifestation of active crustal movement and resurfacing is accompanied by clear evidence of internal differentiation. Unlike Callisto, Ganymede has an iron-rich core and a permanent magnetic field that is strong enough to create its own magnetosphere and auroras.

Europa. The surface of Europa is totally different from that of Ganymede or Callisto, despite the fact that the infrared spectrum of this object indicates that it, too, is covered with ice. There are few impact craters on Europa, indicating that the surface is relatively recent. Some scientists think the surface is so young that significant resurfacing is still taking place on the satellite. This resurfacing evidently consists of the outflow of water from the interior to form an instant frozen ocean.

Models for the differentiated interior suggest the presence of an iron-rich core surrounded by a silicate mantle surmounted by an icy crust some 150 kilometres thick. This moon possesses both induced and intrinsic magnetic fields. Slightly mottled regions on the surface have been found to contain salt deposits, suggesting evaporation of water from a reservoir below the crust. Europa's frozen surface is criss-crossed with dark and bright stripes and curvilinear ridges and grooves. The relief is extremely low, with ridge heights perhaps a few hundred metres at most. Europa thus has the smoothest surface of any solid body examined in the solar system thus far. The major open question is whether there is an ocean of liquid water beneath Europa's ice, warmed by the release of tidal energy in Europa's interior. If such an ocean and source of heat exist, the possible presence of at least microbial life forms must be admitted.

Possible water ocean on Europa

Io. Seen through a telescope from Earth, Io appears reddish orange. Its infrared spectrum shows no evidence of the absorption characteristics of water ice. Scientists expected Io's surface to look different from those of Jupiter's other moons, but the Voyager images revealed a landscape even more unusual than anticipated. Volcanic fissures, instead of impact craters, dot the surface. Nine volcanoes were observed in eruption when the two Voyager spacecraft flew by in 1979, while the closer encounters by Galileo indicated that as many as 300 volcanic vents may be active at a given time. This unprecedented level of activity makes Io the most tectonically active object in the solar system. The surface of the satellite is continually and completely replaced by this volcanism in just a few thousand years.

Various forms (allotropes) of sulfur appear to be responsible for the black, orange, and red areas on the moon's surface, while solid sulfur dioxide is probably the main constituent of the white areas. Sulfur dioxide was detected as a gas near one of the active volcanic plumes by Voyager's infrared spectrometer and was identified as a solid in ultraviolet and infrared spectra obtained from Earth-orbital and ground-based observations. These identifications provide sources for the sulfur and oxygen ions observed in the Jovian magnetosphere and prove that Io's volcanic activity is the source of its torus of particles.

The energy for this volcanic activity requires a special explanation, since radioactive heating is inadequate for a body as small as Io. The favoured explanation is that tides developed by a gravitational tug-of-war over Io between the other Galilean satellites and Jupiter release enough energy to account for the observed volcanism. The interior contains a dense, iron-rich core, which probably produces a magnetic field. The interactions of Io with Jupiter's magnetosphere and ionosphere are so complex, however, that it has been difficult to distinguish the satellite's own field from the current-produced fields in its vicinity.

Cause of Io's volcanism

NASA/JPL/Caltech

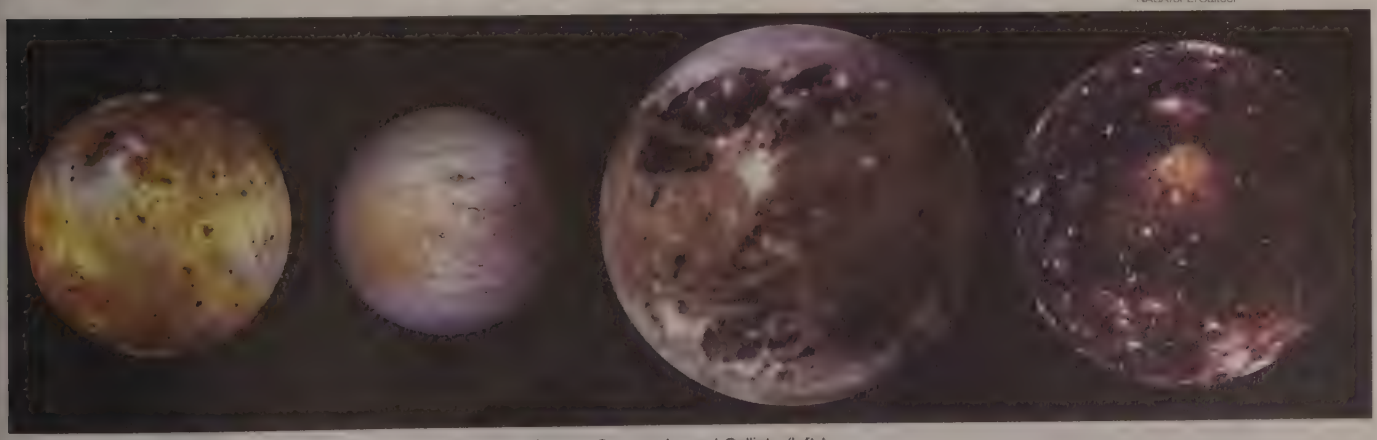


Figure 46: The four Galilean moons of Jupiter—Io, Europa, Ganymede, and Callisto (left to right)—assembled from images made by the Galileo spacecraft in 1996–97. Diameters of the moons are to scale. Colours have been enhanced to highlight surface features.

Other satellites. The only other Jovian moon that was close enough to the trajectories of the Voyager spacecraft to allow surface features to be seen was the tiny, irregularly shaped Amalthea. Like Io, its surface exhibits a reddish colour, which may result from a coating of sulfur compounds released by Io's volcanoes. In addition to providing new images of Amalthea, the Galileo orbiter was able to view the effect of impacts on Thebe and Metis. All three of these inner moons are tidally locked, keeping the same face oriented toward Jupiter. All three are some 30 percent brighter on their leading sides, presumably as a result of impacts by small meteoroids.

Before the turn of the 21st century, the eight outer moons then known had formed two groups: a more distant group in retrograde orbits around Jupiter—Ananke, Carme, Pasiphae, and Sinope—and a closer group in prograde orbits—Leda, Himalia, Lysithea, and Elara. (In the case of these moons, retrograde motion is in the direction opposite to Jupiter's spin and motion around the Sun, which are counterclockwise as viewed from above Jupiter's north pole, whereas prograde, or direct, motion is in the same direction.) With the discovery of a spate of additional outer moons beginning in 1999, this division was recognized as an oversimplification. There are clearly several groups of moons, each group apparently originating from an individual body that was captured by Jupiter and then broke up. The captures could have occurred near the time of Jupiter's formation when the planet was itself surrounded by a nebula that could slow down objects that entered it.

The ring system. As the Pioneer 10 spacecraft sped toward its closest approach to Jupiter in 1974, it detected a sudden decrease in the density of charged particles roughly 125,000 kilometres from Jupiter, just inside the orbit of its innermost moon, Metis. This led to the suggestion that a moon or a ring of material might be orbiting the planet at this distance. The existence of a ring was verified in 1979 by the first Voyager spacecraft when it crossed the planet's equatorial plane, and the second spacecraft recorded additional pictures, including a series taken in the shadow of the planet looking back at the ring material. The complex nature of this ring was elucidated by images obtained with the Galileo spacecraft in 1996–97.

The ring consists of four distinct components: an outer gossamer ring, whose outer radius coincides with the orbital radius of the satellite Thebe (222,000 kilometres); an inner gossamer ring bounded by the orbit of Amalthea (181,000 kilometres); the main ring, extending inward some 6,000 kilometres from the orbits of Adrastea (129,000 kilometres) and Metis (128,000 kilometres); and a halo of particles with a thickness of 25,000 kilometres that extends from the main ring inward to 92,000 kilometres. The ring comprises large numbers of micrometre- and smaller-sized particles that produce strong forward scattering of incident sunlight.

Association of rings and innermost moons

The presence of such small particles requires a source, and the association of the ring boundaries with the four satellites makes the source clear. The ring particles are generated by impacts on these satellites by micrometeoroids, cometary debris, and possibly volcanically produced material from Io. Some of the finest particles are electrically charged and respond to the rocking motion of the Jovian magnetic field as the planet rotates.

THEORIES OF THE ORIGIN OF THE JOVIAN SYSTEM

Explaining the origin of Jupiter and its satellites is part of the problem of explaining the origin of the solar system. Current thinking favours the gradual development of the Sun and planets from a huge cloud of gas and dust containing gravitational instabilities.

Early history of Jupiter. Given the planet's large proportion of hydrogen and its huge mass, it has been traditional to assume that Jupiter formed by condensation from the primordial solar nebula. This hypothesis implies that the elements should all be present on Jupiter in the same proportions that they occur in the Sun. However, the most recent evidence (see Table 13) indicates that the elemental proportions on Jupiter differ from the solar values.

Current models for Jupiter's origin suggest instead that a solid core of about 10 Earth masses formed first as a result of the accretion of icy planetesimals. This core would have developed an atmosphere of its own as the planetesimals released gases during accretion. As the mass of the core increased, it would have become capable of attracting gases from the surrounding solar nebula, thus accumulating the huge hydrogen-helium envelope that constitutes Jupiter's atmosphere and fluid mantle. The accumulating envelope would have mixed with the outgassed atmosphere from the core. Thus, the presently observed enrichment of the most abundant heavy elements in this envelope, compared with solar values (see Table 13), reflects the concentration of such elements in the core. As noted above, the mass spectrometer on the Galileo probe showed that these heavy elements are enriched by the same factor of about three. For this enrichment to include volatile substances like argon and molecular nitrogen requires that the icy planetesimals must have formed at temperatures of 30 K (−400° F, −240° C) or less. Just how this happened remains a puzzle, and its solution may ultimately help explain the presence of giant planets that have been detected very close to their stars in other planetary systems.

Explanation for elemental enrichments in Jovian atmosphere

Early history of the satellites. The inner eight moons of Jupiter are thought to have originated in much the same way as the planet itself. Just as the primordial solar nebula is believed to have broken up into accreting planetesimals, which became the planets, and a central condensation, which became the Sun, the accumulation of material into a protoplanetary cloud at Jupiter's orbit led to the formation of the planet and its inner satellites. The analogy goes further. The high temperature of the forming Jupiter apparently prevented volatile substances from condensing at the distances of the innermost satellites. Hence, Ganymede and Callisto, the most distant of the inner eight moons, represent the volatile-rich outer bodies in this system.

The origin of the outer satellites, with their orbits of high eccentricities and inclinations, is thought to be quite different. They are members of the population of irregular satellites in the solar system and have most likely been captured by Jupiter. Their ultimate origin, however, remains unclear. They may have first formed in the outer nebula of Jupiter, strayed away, and then been recaptured; alternately, they may have formed independently in the solar nebula itself and then been captured. Ongoing studies of these objects and their possible relatives among the Trojan asteroids may provide the answer. (T.C.O.)

Saturn

Saturn, designated ♄ in astronomy, is the second largest planet in mass and size. Its dimensions almost equal those of Jupiter, while its mass is about three times smaller; it has the lowest mean density of any object in the solar system. Both Saturn and Jupiter resemble stellar bodies in that their bulk chemical composition is dominated by hydrogen. However, Saturn's structure and evolutionary history differ significantly from its larger counterpart. Like the other giant planets Jupiter, Uranus, and Neptune, Saturn has an extensive satellite and ring system, which may provide clues to its origin and evolution. Saturn's dense and extended rings, which lie in its equatorial plane, are currently the most impressive in the solar system.

Saturn is the sixth planet in distance from the Sun, with an orbital semimajor axis of 1.427 billion kilometres. It never approaches the Earth more closely than about 1.2 billion kilometres, and thus Earth-based observations of Saturn always show a nearly fully illuminated disk, unlike the Voyager 1 image shown in Figure 47.

PRINCIPAL CHARACTERISTICS

Like most planets, Saturn has a regular orbit with prograde motion around the Sun and a small eccentricity and inclination to the ecliptic. In this regard, it resembles its inner neighbour Jupiter. Unlike Jupiter, however, Saturn has a substantial obliquity, or inclination of its equatorial plane to its orbital plane, of 26.7°. As a result, Saturn's rings are presented to Earth-based observers at opening angles ranging from 0° (edge on) to nearly 30°.

Saturn has no single rotation period. Cloud motions in its massive upper atmosphere can be used to trace out a

Figure 47: View of Saturn from Voyager 1 on Nov. 16, 1980, four days after its closest approach, at a distance of 5.3 million kilometres.
B.A. Smith/National Space Science Data Center

variety of rotation periods, with periods as short as about 10 hours, 10 minutes near the equator and increasing with some oscillation to about 30 minutes longer at latitudes higher than 40°. The rotation period of Saturn's deep interior can be determined from the rotation period of the magnetic field, which is presumed to be rooted in a metallic outer core. Measurement of the field's rotation is difficult because the field is highly axisymmetric. Small irregularities in the field appear to be related to periodic radio outbursts in the magnetosphere with a period of 10 hours, 39.4 minutes, which is taken to be the magnetic field rotation period. There are also radio bursts with periods of about 10 hours, 10 minutes, which originate with lightning in Saturn's atmosphere.

Magnetic field rotation

The equatorial diameter of Saturn, 120,536 kilometres, is measured with respect to the one-bar pressure level in its atmosphere, for Saturn has no solid surface in its outer layers. Saturn is the most oblate (flattened at the poles) of all the planets in the solar system, with a polar diameter (at one bar) of 108,728 kilometres, 10 percent smaller than the equatorial diameter. Correspondingly, the equatorial gravity of the planet, 8.96 metres per second squared (m/s^2), is only 74 percent of the polar gravity, 12.14 m/s^2. The mass of Saturn is 5.685×10^{26} kilograms, or 95.13 times the mass of the Earth, while its volume is 766 times the volume of the Earth. Saturn's mean density is 0.69 gram per cubic centimetre. The escape velocity from the one-bar level is high, 36 kilometres per second, and thus there has been no significant escape of gas from the planet since its formation. See Table 15 for some characteristics of Saturn.

THE ATMOSPHERE

Saturn's atmosphere is 91 percent hydrogen by mass and is thus the most hydrogen-rich atmosphere in the solar system. Helium, which is measured indirectly, comprises another 6 percent and is less abundant relative to hydrogen as compared with a gas of solar composition. If hydrogen, helium, and other elements were present in the same proportions as in the Sun's atmosphere, Saturn's atmosphere would be about 71 percent hydrogen and 28 percent helium by mass.

The remaining major molecules that have been observed in Saturn's atmosphere are methane (CH_4) and ammonia (NH_3), which are a factor of two to five times more abundant relative to hydrogen than in a gas of solar composition. Hydrogen sulfide (H_2S) and water (H_2O) are expected to be major constituents of the deeper atmosphere but have not yet been detected. Minor molecules that have been spectroscopically detected include phosphine (PH_3), carbon monoxide (CO), and germane (GeH_4); such molecules would not be present in detectable amounts in a hydrogen-rich atmosphere in chemical equilibrium. They may therefore be disequilibrium products of reactions at high pressure and temperature in Saturn's deep atmosphere well below the observable clouds. A number of disequilibrium hydrocarbons are observed in Saturn's stratosphere: acetylene (C_2H_2), ethane (C_2H_6), and, possibly, propane (C_3H_8) and methyl acetylene (C_3H_4). All of the latter may be produced by photochemical effects from solar radiation or by energetic particle bombardment.

Analysis of the refraction of starlight and radio waves has provided information on the distribution of temperature in Saturn's atmosphere from pressures of one-millionth bar to 1.3 bar. At pressures below 1 millibar the atmosphere is roughly isothermal at about 140–150 K. A stratosphere, where temperatures steadily decline with increasing pressure, extends from 1 to 60 millibars, where the coldest temperature in Saturn's atmosphere (82 K) occurs. At higher pressures the temperature increases once again in the troposphere, following the so-called adiabatic lapse rate. This region is analogous to the Earth's troposphere, in which the increase of temperature with pressure follows the thermodynamic relation for compression of a gas

Temperature distribution

Table 15: Planetary Data for Saturn

Distance from the Sun	mean 1,427,000,000 km maximum 1,507,000,000 km minimum 1,347,000,000 km
Eccentricity of orbit	0.056
Inclination of orbit to ecliptic	2.5
Sidereal period of revolution	29.46 years
Rotation period (magnetic field)	10 h 39.4 min
Mean synodic period	378.09 Earth days
Obliquity	26.7°
Diameter at 1 bar (equatorial)	120,536 km
Diameter at 1 bar (polar)	108,728 km
Mass	5.685×10^{26} kg
Volume	766 × Earth volume
Average density	0.69 g/cm³
Equatorial gravity	8.96 m/s²
Polar gravity	12.14 m/s²

without gain or loss of heat. Saturn's tropospheric lapse rate is significantly affected by the quantum mechanics of hydrogen molecules at low temperatures. The temperature is 135 K at a pressure of 1 bar and continues to increase at higher pressures following the adiabatic relation.

The critical point of hydrogen (the highest temperature and pressure at which liquid and gas phases can exist in equilibrium) occurs at 33 K and 13 bars. Since Saturn's atmosphere is everywhere at a temperature of 82 K or higher, the hydrogen behaves as a supercritical liquid as it is compressed without gain or loss of heat. Thus, there is no distinct interface between the higher atmosphere where the hydrogen behaves predominantly as a gas and the deeper atmosphere where it resembles a liquid. Saturn's troposphere does not terminate on any solid surface, but it apparently extends tens of thousands of kilometres below the visible clouds, reaching temperatures of thousands of kelvins and pressures in excess of one million bars.

Like other giant planets, Saturn's atmospheric circulation is dominated by zonal (east-west) flow. When referenced to the rotation of the magnetic field, virtually all the flow is to the east—*i.e.*, in the direction of rotation. A particularly active eastward flow is observed in the equatorial zone at latitudes below 20°, with a maximum zonal velocity of almost 0.5 kilometres per second. This equatorial jet is analogous to one on Jupiter but extends twice as wide in latitude and moves four times faster.

Zonal jets

The zonal flows are remarkably symmetric about Saturn's equator; that is, each jet at a given northern latitude has a counterpart at a similar southern latitude. Strong eastward jets (relative velocities in excess of 100 metres per second) are seen at 46° north and south and at about 60° north and south. Westward jets, which are nearly stationary in the magnetic field's frame, are seen at 40°, 55°, and 70° north and south. Earth-based observations of Saturn's clouds over many years agree with the detailed spacecraft observations of the jets and thus corroborate their stability over time.

The north-south symmetry suggests that the zonal flows may be connected in some fashion deep within the interior. Theoretical investigations have shown that differential rotation of a deep-convecting fluid planet will tend to occur along cylinders aligned about the mean rotation axis. Saturn's atmosphere may display a series of coaxial cylinders, each rotating at a unique rate, which give rise to the zonal jets at the surface. The continuity of the cylinders may be broken at a point where they intersect a major discontinuity within Saturn, such as a core (see Figure 48.)

The atmosphere of Saturn shows many smaller-scale time-variable features similar to those found in Jupiter, such as red, brown, and white spots, bands, eddies, and vortices. The atmosphere generally has a much blander appearance than Jupiter's, however, and is less active on a small scale. A spectacular exception occurred during September–November 1990, when a large white spot appeared near the equator, expanded to a size exceeding 20,000 kilometres, and eventually spread around the equator before fading.

From A.P. Ingersoll *et al.*, "Structure and Dynamics of Saturn's Atmosphere," after F.H. Busse, "A Simple Model of Convection in the Jovian Atmosphere," *Icarus*, vol. 29 (1976), in T. Gehrels and M.S. Matthews (eds.) *Saturn* (1984), The University of Arizona Press, Tucson

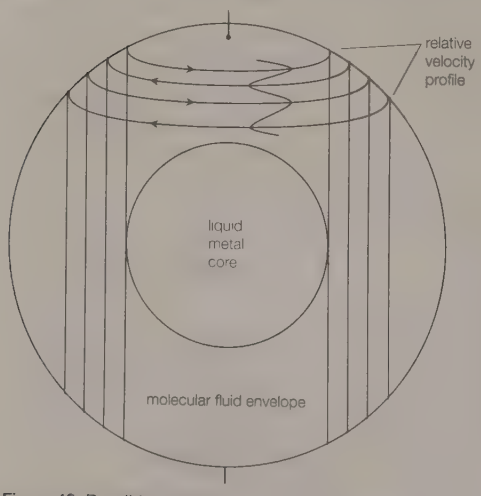


Figure 48: Possible cylindrical zonal flow in Saturn's interior.

The "surface" of Saturn that is seen through telescopes and in spacecraft images is actually a complex layer of clouds formed from molecules of minor species that condense in the hydrogen-rich atmosphere. Although aerosol particles formed from photochemical reactions are seen high in the atmosphere at pressures of 20–70 millibars, the main clouds commence at pressures exceeding 400 millibars, with the highest cloud deck expected to be formed of solid ammonia crystals. The base of the ammonia cloud deck is predicted to occur at a pressure of about 1.7 bars, where the ammonia crystals dissolve into the hydrogen gas and disappear abruptly. Nearly all information about deeper cloud layers has been obtained indirectly by constructing chemical models of the behaviour of compounds expected to be present in a gas of near solar composition following the temperature-pressure profile of Saturn's atmosphere. The bases of successively deeper cloud layers occur at 4.7 bars (ammonium hydrosulfide [NH_4SH] crystals) and at 10.9 bars (water-ice crystals with aqueous ammonia droplets). The actual clouds of Saturn display various shades of yellow, brown, and red, whereas all of the above clouds are colourless in the pure state. Thus, the observed shades are apparently produced by chemical impurities (phosphorus-bearing molecules are a prime candidate).

Cloud layers

INTERIOR STRUCTURE AND COMPOSITION

The low mean density of Saturn is direct evidence of the preponderance of hydrogen in its bulk composition. Under Saturnian conditions, hydrogen behaves as a liquid rather than a gas at pressures exceeding about one kilobar (corresponding to a depth of 1,000 kilometres below the clouds). At this depth the temperature is roughly 1,000 K, much higher than the critical temperature of hydrogen, and thus there is no identifiable interface at which the hydrogen layers are gaseous above and liquid below to distinguish between Saturn's atmosphere and interior. Even as a liquid, hydrogen is a highly compressible material, and a pressure in excess of one megabar is required to attain a density equal to the mean density of Saturn. Such pressure is achieved at a depth of 20,000 kilometres below the clouds.

Information about the interior structure of Saturn is obtained from studying its gravitational field, which is not spherically symmetric. The planet's rapid rotation and low mean density lead to distortion of its physical shape and the shape of its gravitational field, which can be measured precisely from the motion of spacecraft and eccentric ringlets. The degree of distortion from spherical symmetry is directly related to the relative amounts of mass concentrated in Saturn's central regions as opposed to its envelope. Such an analysis shows that Saturn is substantially more centrally condensed than Jupiter and therefore contains a significantly larger amount of material denser than hydrogen near its centre. Saturn's central regions contain about 50 percent hydrogen by mass, while Jupiter's contain approximately 67 percent hydrogen.

At a pressure of roughly two megabars and a temperature of about 6,000 K, the hydrogen is predicted to undergo a major phase transition to so-called liquid metallic hydrogen, which resembles a molten alkali metal such as lithium. This transition occurs at a radius about halfway between Saturn's atmosphere and centre. Evidence from the planet's gravitational field shows that the central metallic region is considerably denser than would be the case for pure hydrogen with solar proportions of helium. It is likely that the depletion of helium in Saturn's atmosphere is compensated by an excess of helium in the deeper metallic region, partially accounting for the increased density. A substantial quantity (perhaps nearly 30 Earth masses) of material denser than both hydrogen and helium may also be present in Saturn, but its precise distribution cannot be determined from available data. A rock and ice mixture

Liquid metallic hydrogen

of approximately 10–20 Earth masses is likely to be concentrated in a dense central core.

On average, Saturn absorbs 11×10^{16} watts of power from the Sun, while it radiates 20×10^{16} watts into space, primarily at infrared wavelengths between 20 and 100 micrometres. The difference between these numbers represents Saturn's present internal power, which must be derived from interior heat-generating processes. The specific internal power, which is the internal power per unit mass, is 1.5×10^{-10} watts per kilogram, which may be compared with the corresponding value for the Sun, 1.9×10^{-4} watts per kilogram, and for Jupiter, 1.7×10^{-10} watts per kilogram.

Although Saturn's specific internal power is similar to Jupiter's, it is evidently derived at least partially from a different source. A calculation of thermal evolution shows that Saturn could have originated with the gravitational collapse of gaseous hydrogen and helium from the original solar nebula onto a massive ice-rich core of perhaps 10 to 20 Earth masses. The core may have had a composition similar to that of the present icy Saturnian satellites. Jupiter may have undergone a similar origin, but with a greater amount of gas captured. The gas was heated to high temperatures (several tens of thousands of degrees kelvin) in the course of the capture. Jupiter's present internal power can then be understood as residual cooling of an initially hot planet over the age of the solar system, some 4.6 billion years, a mechanism similar to one once proposed (unsuccessfully) to explain the Sun's internal power. For Saturn, application of such a mechanism predicts a value for the cooling time that is too low by about a factor of two. That is to say, if the planet is assumed to be initially formed at a high temperature, the internal power drops below the present observed value after only 2.6 billion years. It has been theorized that the onset of hydrogen-helium immiscibility and thus the gradual sinking of helium liberates additional gravitational energy. As the helium separates from hydrogen in the metallic phase of hydrogen and "rains" into deeper levels, potential energy is converted into kinetic energy of helium droplet motion. This motion is then damped by friction and converted into heat, which is radiated into space, thus prolonging the duration of Saturn's internal power. This process is not believed to occur in Jupiter, which has a warmer interior and thus more helium miscibility. Detection of a substantial depletion of helium in Saturn's atmosphere by the Voyager spacecraft was taken as a vindication of the theory, details of which remain controversial.

Helium rain

MAGNETIC FIELD AND MAGNETOSPHERE

Saturn's magnetic field resembles that of a simple dipole or bar magnet with the axis of symmetry closely aligned (to within one degree) with Saturn's rotation axis and the centre of the equivalent dipole at the centre of the planet. The polarity of the field is opposite to that of the Earth's field; *i.e.*, the field lines emerge in Saturn's northern hemisphere and reenter the planet in the southern hemisphere. There are measurable deviations from a simple dipole field, which manifest themselves both in a north-south asymmetry and in a slightly higher polar surface field than would be predicted in a pure dipole model. The maximum polar surface field is 0.8 gauss (north) and 0.7 gauss (south), very similar to the Earth's polar surface field, while the equatorial surface field is 0.2 gauss.

The calculated electrical conductivity of Saturn's liquid metallic-hydrogen core is approximately 10^5 mhos per centimetre, about the same as that of lithium at one atmosphere pressure and a temperature just above its melting point. If slow circulation currents are present, as would be expected with the flow of heat to the surface accompanied by gravitational settling of denser components, sufficient dynamo action is expected to produce the observed magnetic field. Saturn's field is thus produced by essentially the same mechanism as produces the Earth's field. The deep field, in the vicinity of the dynamo region near the core, may be quite irregular. Theories hypothesize that magnetic field lines are made more axisymmetric before they reach the surface by passing through a nonconvecting, electrically conducting region that is rotating with respect to the field lines. Saturn's striking atmospheric differential rotation may be related to the action of much deeper currents involving the conducting core.

Generation of magnetic field

Saturn's magnetosphere (the region of space dominated by Saturn's magnetic field rather than interplanetary magnetic fields) extends to a distance of about 20 Saturn radii from the centre of the planet on the sunward side but with substantial fluctuation due to variations in the dynamic pressure from the solar wind. On the antisunward side the magnetosphere is drawn out into a long magnetotail, which extends to much greater distances. Saturn's satellites Titan and Hyperion orbit at distances close to the minimum magnetospheric dimensions and occasionally cross the boundary. As a consequence, charged particles from Titan's upper atmosphere may interact with the local magnetic field lines.

Saturn's inner magnetosphere, like the magnetospheres of the Earth and Jupiter, has a stable population of energetic protons (those having energies greater than tens of millions of electronvolts) spiraling along magnetic field lines. Unlike the magnetospheres of the Earth and Jupiter, however, this population is substantially altered by absorption of the energetic particles onto the surfaces of solid bodies orbiting within the field lines. "Holes" are seen in the particle populations on field lines that thread the rings and the orbits of satellites within the magnetosphere.

THE SATELLITES AND RINGS

Satellite system. Saturn possesses an extensive system of satellites, and all but the outermost have prograde, low-inclination and low-eccentricity orbits with respect to the planet. (Data on the Saturnian satellites are summarized in Table 16.) A small satellite (S18) with dimensions on the order of 10 kilometres has been observed within the ring system; its presence within the Encke gap (an area of decreased brightness in the A ring; see below *Ring system*) was deduced from its dynamical effects on the gap, and more such ring moons doubtlessly remain to be discovered. Indeed, the ring system itself consists of myriads of much smaller satellites.

The orbital and rotational dynamics of Saturn's satellites show complexities unique to this system. The small satellites S10, S11, and S17 orbit near the outer edge of the main ring system and should dynamically interact with it by receiving angular momentum from ring particles through collective gravitational interactions. The effects of this process are to reduce the spreading of the rings caused by inelastic collisions between ring particles and to drive these satellites to larger orbital radii. Because of the small size of the satellites, it is difficult to find a mechanism by which this process could have endured over the age of the solar system without driving the satellites far beyond their current positions. The sharpness of the outer edge of the ring system and the present orbits of the inner satellites are puzzling, and such features may imply that the current ring system is much younger than Saturn itself.

The satellites S15 and S16 are classical shepherd satellites, orbiting on either side of the F ring and confining its particles to a narrow band. The inner shepherd (S16) transmits angular momentum to the ring, pushing the ring outward and itself inward, while the outer shepherd (S15) receives angular momentum from the ring, pushing the ring inward and itself outward. The co-orbital satellites Janus and Epimetheus (S10 and S11) share the same average orbit, interacting with each other every few years in such a way that one transmits angular momentum to the other, which forces the latter into a slightly higher orbit and the former into a slightly lower orbit. At subsequent closest approach, the process repeats in the opposite direction. The satellites Tethys (S3) and Dione (S4) also have co-orbital satellites, but since Tethys and Dione are much more massive than their co-orbiters, there is no significant exchange of angular momentum. Instead, Tethys' co-orbiters S13 and S14 are located at the stable Lagrange points on Tethys' orbit, leading and following Tethys by 60°, in a manner precisely analogous to the Trojan asteroids in Jupiter's orbit. Dione possesses only a single Trojan-like companion, S12, which leads it by 60° on average.

Co-orbital satellites

Table 16: Satellite Data for Saturn*

satellite	a	P (days)	e	i (°)	radius (km)	mass (10^{23} g)	density (g/cm^3)
S18 1981S13	2.227	0.583	—	—	~10	—	—
S17 Atlas	2.276	0.602	0.002	0.3	20×?×10	—	—
S16 1980S27	2.310	0.613	0.004	0.0	70×50×37	—	—
S15 1980S26	2.349	0.629	0.004	0.1	55×45×33	—	—
S10 Janus	2.51†	0.69†	—†	—†	110×95×80	—	—
S11 Epimetheus	2.51†	0.69†	—†	—†	70×58×50	—	—
S1 Mimas	3.08	0.94	0.020	1.5	197±3	0.46±0.05	1.4±0.2
S2 Enceladus	3.95	1.37	0.004	0.0	251±5	0.8±0.3	1.2±0.5
S3 Tethys	4.88	1.89	0.000	1.1	530±10	7.6±0.9	1.2±0.1
S13 Telesto	4.88	1.89	—	—	15×10×8	—	—
S14 Calypso	4.88	1.89	—	—	12×11×11	—	—
S4 Dione	6.26	2.74	0.002	0.0	560±5	10.5±0.3	1.4±0.1
S12 1980S6	6.26	2.74	0.005	0.2	17×16×15	—	—
S5 Rhea	8.73	4.52	0.001	0.4	765±5	24.9±1.5	1.3±0.1
S6 Titan	20.3	15.95	0.029	0.3	2,575±0.5	1,345.7±0.3	1.88
S7 Hyperion	24.6	21.28	0.104	0.4	205×130×110	—	—
S8 Iapetus	59	79.3	0.028	14.7	730±10	18.8±1.2	1.2±0.1
S9 Phoebe	215	550	0.163	150	110±10	—	—

*a is the orbital semimajor axis in units of Saturn's equatorial radius, P is the sidereal orbital period, e is the orbital eccentricity, and i is the inclination of the orbit plane to Saturn's equatorial plane. †Parameters are variable in these co-orbital satellites due to orbital exchange.
Source: Adapted from D. Morrison, T.V. Johnson, E.M. Shoemaker, L.A. Soderblom, P.T. Thomas, J. Veverka, and B.A. Smith, "Satellites of Saturn: Geological Perspective," in T. Gehrels and M.S. Matthews, eds., *Saturn*, The University of Arizona Press, copyright © 1984.

The satellite pairs Mimas-Tethys, Enceladus-Dione, and Titan-Hyperion are in stable dynamical resonances, in the sense that they interact gravitationally in a periodic fashion so as to preserve the regularity. The ratio of the orbital periods of a satellite pair in resonance is approximately equal to a ratio of small whole numbers. The orbital periods of Titan and Hyperion are in the ratio 3:4, such that conjunction of Titan and Hyperion always occurs at Hyperion's apoapse (farthest point of its elliptical orbit from Saturn). Since the much larger body Titan always exerts a maximum gravitational perturbation at the same points on Hyperion's orbit, Hyperion is forced into a relatively large eccentricity. Analogously, Enceladus and Dione have orbital periods in the ratio 1:2, as is also the case for Mimas and Tethys. Resonances may be important in the structure of the system of the concerned satellites because they can force orbital eccentricities to relatively large values. Ordinarily, tidal interactions between satellites and Saturn tend to circularize the satellite orbits as well as to force the satellites into synchronous rotation. Once such a rotation state has been established, tides are stationary in the satellite's frame and do not cause energy dissipation. However, eccentricity forced by resonance causes time-variable tides on satellites, with accompanying energy dissipation and heating of the satellite's interior. Although calculations indicate that present tides on Saturn's satellites are not particularly significant as a heating mechanism, this may not have been true in the past. A tenuous and diffuse outer ring of Saturn, the so-called E ring, is associated with the orbit of Enceladus and may consist of material supplied by episodes of volcanism on Enceladus.

Although tidal friction ordinarily forces satellites into a state of synchronous rotation, Hyperion appears to be a spectacular exception to this rule. Because of its large orbital eccentricity and highly unspherical shape, there is a complicated interaction between Hyperion's spin and orbital angular momentum leading to a chaotic feedback process. Although Hyperion was observed from the Voyager spacecraft to be rotating with a nonsynchronous period of about 13 days, chaos theory shows that it is actually tumbling in an essentially unpredictable manner. As Mercury is the only object in the solar system known to be captured into a resonance with a ratio of rotational period to orbital period other than the usual 1:1 (Mercury's ratio is 2:3), so Hyperion is the only object known to be in chaotic rotation.

Titan

Titan is the only known satellite in the solar system with a dense atmosphere. The atmosphere was first detected spectroscopically in 1944, by Gerard P. Kuiper, who found evidence of methane absorption. However, studies of refraction of radio waves in the atmosphere carried out by the Voyager 1 spacecraft have shown that methane is a minor constituent of the atmosphere, comprising only 2 to 10 percent by number, and that spectroscopically inactive molecules predominate. Comparison of infrared and radio data shows that the mean molecular weight of the atmosphere is 28.6 atomic mass units. Thus, the most plausible major constituent is nitrogen (N_2; mean molecular weight 28), although some argon could also be present (mean molecular weight 36). The atmosphere of Titan is similar to that of the Earth in that it consists predominantly of nitrogen gas and has a surface pressure of 1.5 bars. However, Titan's atmosphere is much colder than the Earth's, with a surface temperature of 94 K, and contains no free oxygen. Titan visually appears as a uniform brownish globe; its surface is permanently veiled by dense clouds of uncertain composition, extending to altitudes of hundreds of kilometres.

A small troposphere extends from Titan's surface to an altitude of 42 kilometres, where a minimum temperature of 71 K is reached. Apparently temperatures are always above the condensation point of nitrogen, so that nitrogen clouds are not present. Methane clouds may be present in this region if methane has a mixing ratio exceeding 1.5 percent, but a liquid methane ocean at Titan's surface would require a mixing ratio of 12 percent and is not expected. The existence of an ocean of ethane (C_2H_6) has been suggested, but radar echoes from Titan reveal a mostly solid surface.

Titan's stratosphere extends from 50 to 200 kilometres in altitude, with temperatures steadily increasing with altitude to maximum values between 160 and 180 K. Studies of the refraction of starlight in Titan's upper atmosphere show that temperatures remain in this range up to altitudes of 450 kilometres, and spacecraft observations of the transmission of solar ultraviolet light give similar values at even higher altitudes. Many carbon-bearing molecules produced by photochemical processes in Titan's high atmosphere have been detected spectroscopically. These include carbon monoxide (CO), ethane (C_2H_6), propane (C_3H_8), acetylene (C_2H_2), ethylene (C_2H_4), hydrogen cyanide (HCN), methyl acetylene (C_3H_4), diacetylene (C_4H_2), cyanoacetylene (HC_3N), cyanogen (C_2N_2), and carbon dioxide (CO_2), all detected in trace amounts.

Particulates that absorb solar radiation are extraordinarily pervasive throughout Titan's atmosphere, attaining substantial tangential optical depth even at altitudes of 300 kilometres and gas pressures substantially below one millibar. Their typical sizes probably lie in the range of 0.1 micrometre. There is evidence that they grow to a substantially higher density in Titan's summer hemisphere, suggesting that they are a form of natural photochemical "smog." Solar heating of the particle layers creates a temperature inversion layer in Titan's stratosphere, preventing dissipation of the smog layer by convection.

Saturn's other satellites are much smaller than Titan and possess no detectable atmospheres. However, their low mean densities, as well as the spectroscopy of their

surface solids, indicate that they are rich in ice, probably mostly water ice, with perhaps some solids of more volatile species (possibly ammonia). Under the low solar equilibrium temperatures prevailing at Saturn, the ice behaves mechanically like rocky material in the inner solar system, and the surfaces of the satellites bear a superficial resemblance to the cratered rocky surface of the Moon but with many important differences. Images of satellites were obtained by the Voyager 1 and 2 spacecraft, but they vary in resolution and surface coverage and are in many cases far from complete.

Mimas reveals a heavily cratered surface similar to the lunar highlands, but it also possesses one of the largest impact structures (in relation to the satellite's size) in the solar system. The crater Herschel (named in honour of the 19th-century British astronomer William Herschel), situated on Mimas' leading hemisphere, is 130 kilometres in diameter (one-third of the diameter of Mimas), is roughly 10 kilometres deep, and has outer walls about 5 kilometres high.

Enceladus

Enceladus, which is suspected of possible volcanism, has a highly reflecting surface, with a normal reflectance exceeding that of newly fallen snow (see Figure 49). Few large craters are seen, while extensive ridged plains and crater-free areas give convincing evidence of fairly recent internal activity, possibly within the last 100 million years. Particles in the E ring centred on Enceladus' orbit have sizes in the range of one micrometre and could persist for only a few thousand years. Thus, an Enceladus event that produced them may have occurred that recently.

Tethys, although larger than Enceladus, shows little evidence of internal activity. Its heavily cratered surface appears to be quite old, although it displays subtle features indicative of creep or viscous flow in its icy lithosphere. Dione and Rhea exhibit heavily cratered surfaces similar to the lunar highlands, but with bright patches that may be freshly exposed ice. Although Dione is smaller than Rhea, it has more evidence of recent internal activity, including resurfaced plains and fracture systems.

The surface of Iapetus has not been studied in detail, but it shows a large asymmetry in reflectivity between its leading and trailing hemispheres. The leading hemisphere is remarkably dark, with the darkest material concentrated at the apex of orbital motion. The composition of the dark material is not known with certainty. The trailing hemisphere is heavily cratered, highly reflecting, and appears to be icy in composition. The low mean density of Iapetus suggests that the satellite as a whole is mostly ice.

Ring system. Saturn's ring system ranks among the most spectacular phenomena in the solar system. With a diameter of 270,000 kilometres, the system is an enormous object, yet its thickness does not exceed 100 metres, and its total mass comprises only about 3×10^{22} grams, similar to the mass of Mimas.

Like the rings of the other giant planets Jupiter, Uranus, and Neptune, those of Saturn lie for the most part within the classical Roche limit. This limit, which for the idealized case is at 2.44 Saturn radii, represents the closest distance at which a small satellite can approach a massive primary before it is torn apart by tidal forces. Conversely, small bodies within the Roche limit are prevented from aggregating into larger objects by tidal forces. The limit applies only to objects held together by gravitational attraction and thus does not restrict the stability of relatively small bodies for which molecular cohesion is important. Thus, small moons with sizes in the range of tens of kilometres or less can persist indefinitely within the Roche limit.

Particle size distribution

Although the individual particles that make up Saturn's rings cannot be seen directly, their size distribution can be deduced from their effect on the scattering of signals propagated through the rings from spacecraft and stars. This analysis shows a broad and continuous spectrum of particle sizes, ranging from centimetres to several metres, with larger objects being significantly fewer in number than smaller ones. This spectrum is consistent with the distribution that might be expected from repeated collision and shattering of initially larger objects. In some parts of the rings, where collisions are apparently more frequent, even smaller (dust-sized) grains are present, but these have short lifetimes owing to a variety of loss mechanisms. Clouds of the smaller grains apparently acquire electrical charges, interact with the magnetic field, and manifest themselves in the form of moving, wedge-shaped spokes that extend radially. Much larger ring moons—those of several kilometres—may exist within the rings but are apparently quite rare. Spectra of sunlight reflected from

B A Smith/National Space Science Data Center

Figure 49: View of Enceladus from Voyager 2, showing crater-free portions of the surface, possibly indicative of resurfacing by liquid water from the interior.

the rings show absorption by water ice, and the rings are highly reflective to visual light. Thus, it is conceivable that the rings were produced by the disruption of a satellite of the size and composition of Mimas.

The ring system shows structures on many scales, ranging from the broad divisions into the classical A, B, and C rings down to a myriad of individual ringlets with radial scales on the order of kilometres. The structures have provided a fertile field for investigating gravitational resonances and the collective effects of many small particles orbiting in close proximity. Although many of the structures can be understood theoretically, a large number remain enigmatic, and a complete synthesis of the system is still lacking. Because the Saturn ring system may be an analogue of the original disk-shaped system of particles out of which the solid planets formed, an understanding of its dynamics and evolution has implications for the origin of the solar system itself.

The structure of the rings is broadly described by the optical depth, τ, as a function of radial distance from the centre of Saturn. The optical depth, which is a measure of the average density of the rings, is defined as the natural logarithm of the ratio of the incident intensity of light of a specified wavelength to the emergent intensity, where the light is assumed to propagate in a direction perpendicular to the ring plane. Radio signals with wavelengths of several centimetres and greater are largely unaffected by the smallest ring particles and thus encounter smaller optical depths than signals with wavelengths in the visible region of the electromagnetic spectrum and shorter.

Ring locations and optical depths

The B ring is the thickest and broadest of the rings, extending from 1.52 to 1.95 Saturn radii, with optical depths between 1.2 and 1.8. It is separated from the outer major ring, the A ring, by the Cassini division. The Cassini division (1.95 to 2.02 Saturn radii) is not devoid of particles but exhibits complicated variations in τ, with an average value of 0.12. The A ring extends from 2.02 to 2.27 Saturn radii, with a τ value of about 0.7 to 0.6. Interior to the B ring lies the C ring (sometimes known as the Crepe ring), at 1.23 to 1.52 Saturn radii, with optical depths about 0.1. Interior to the C ring lies the extremely tenuous D ring (1.11 to 1.23 Saturn radii), visible only in reflected light. Exterior to the A ring lies the narrow, shepherded F ring at 2.33 Saturn radii. The tenuous G ring, at 2.8 Saturn radii, was originally detected by its influence on charged particles in Saturn's magnetosphere and is faintly discernible in Voyager images. The E ring extends from 3 to 8 Saturn radii.

Numerous gaps are seen in the distribution of optical depth in the major ring regions. Some of the major gaps have been named after famous astronomers who were associated with studies of Saturn. In addition to the Cassini division, they include the Maxwell gap (1.45 Saturn radii), the Huygens gap (1.95 Saturn radii), the Encke gap (2.21 Saturn radii), and the Keeler gap (2.26 Saturn radii). Of the latter four gaps, only the Encke gap was known prior to the existence of spacecraft images of Saturn.

Formation of gaps

Particles can be cleared from a region to form a gap by the gravitational effects of a moon about 10 kilometres in size orbiting within the gap region; such a moon (S18) was found within the Encke gap. The corresponding moon within the Keeler gap has not yet been found but is believed to exist. Gaps can also be cleared in ring regions that are in orbital resonance with satellites whose orbits are substantially interior or exterior to the rings. The condition for resonance is that orbital periods of the satellite and ring particle be in the ratio l:$(m-1)$, where l and m are integers. When this condition is satisfied, the satellite (if external to the ring) receives angular momentum from the resonant ring particles, launching a tightly wound spiral density wave in the ring and ultimately clearing a gap if the resonance is strong enough. The outer edge of the B ring (inner edge of the Cassini division) is in 2:1 resonance with Mimas and shows the two-lobed excursions in radius predicted by such a resonance. Similarly, the outer edge of the A ring, and of the main ring system itself, is in 7:6 resonance with the co-orbital satellites Janus and Epimetheus and is scalloped with seven lobes. Effects of other resonances with various orbital frequencies of external satellites are seen throughout the ring system, but many similar features have no such explanation.

HISTORY OF OBSERVATION

Saturn is easily visible to the naked eye as a point of light and has thus been known to man since prehistoric times. Its motions were systematically studied by various early cultures, and, as the farthest of the planets known to pre-telescopic astronomers, it was seen to be the slowest-moving. Its modern name comes to us from the Roman god of agriculture, known to the Greeks as Cronus, the father of Zeus.

The ring system was discovered by Galileo in 1610, when he first observed Saturn with a primitive telescope. However, Galileo's instrument produced images that were insufficiently clear for him to discern the true geometry of the system. Rather, he reported that Saturn was a triple planet: a central globe with smaller bodies nearly in contact on either side. In 1612 the image became a single globe because the rings essentially disappeared during the Earth's ring-plane passage. Later observations by Galileo showed that the curious appendages had reappeared. However, he apparently never deduced that the appendages were in fact a flat ring encircling the planet.

The Dutch scientist Christiaan Huygens began studying Saturn with an improved telescope in 1655, and he eventually correctly deduced that the planet was encircled by a flat ring and that the ring plane was inclined substantially to Saturn's orbital plane. He believed that the ring was solid, however, and that it had a substantial thickness. Cassini's discovery of a gap between the rings some years later cast doubt on the possibility of a solid ring, and Pierre-Simon Laplace of France published a theory in 1789 that the rings were made up of many smaller components. In 1857, James Clerk Maxwell demonstrated mathematically that the rings could be stable only if they were made up of a very large number of small particles.

Even under the optimal conditions attainable with an Earth-based telescope, features smaller than a few thousand kilometres on Saturn cannot be resolved. Thus, the great detail exhibited in the rings and atmosphere was largely unknown prior to spacecraft observations. Even the division reported in 1837 by Johann F. Encke of Germany was considered dubious until it was confirmed in 1978 by Harold Reitsema, who used measurements of an Iapetus eclipse by the rings to improve on normal Earth-based resolution.

Modern research on Saturn relies on special techniques used with Earth-based telescopes and space probes. Infrared spectroscopy of the rings, atmosphere, and satellites has yielded considerable information about their composition and thermal balance. Spatial resolution of the rings and atmospheric structures on the scale of kilometres is obtained by observing signals from stars that pass behind the planet as seen from the Earth, as occurred in 1989. In 1990 the "great white spot" was successfully observed with the Hubble Space Telescope from Earth orbit.

The greatest advances in knowledge of Saturn have come from three space probes. In 1973 and 1974, the two spacecraft Pioneer 10 and 11 encountered Jupiter, their nominal objective. Pioneer 10 then proceeded on an orbit into interstellar space. However, it was possible to use the Jupiter encounter to send Pioneer 11 on to Saturn, although this was not originally an objective of the mission. The retargeting was successful, and on Sept. 1, 1979, Pioneer 11 became the first man-made object to reach Saturn, passing through the ring plane at a distance of only 38,000 kilometres from the A ring and passing within 21,000 kilometres of Saturn's atmosphere. The Voyager 1 and 2 spacecraft, which carried much more elaborate imaging equipment, were specifically designed for scientific objectives at Saturn and encountered the planet on Nov. 12, 1980, and Aug. 25, 1981, respectively, after first encountering Jupiter. Intensive further study of the Saturn system will involve the use of a spacecraft specifically designed to orbit the planet and encounter its satellites repeatedly so as to secure more detailed information about the composition, geology, meteorology, and dynamics of the many bodies in the region. (W.B.H.)

Voyager encounters

Uranus

Uranus, symbol ♅ in astronomy, is the seventh planet in order of distance from the Sun. Its low density (1.285 grams per cubic centimetre) and large size (radius four times that of the Earth) place it among the four giant planets, which have no solid surfaces and are composed primarily of hydrogen, helium, water, and other volatile compounds. Absorption of red light by methane gas gives the planet a blue-green colour (see Figure 50). Uranus spins on its side; its rotation axis is tipped at an angle of 98° relative to its orbit axis. In addition, the magnetic field is tipped at an angle of 59° relative to the rotation axis. Uranus has 15 known satellites, ranging up to 789 kilometres in radius, and 10 narrow rings. Its mean distance from the Sun is 2,870.99 million kilometres, and its mean distance from the Earth at closest approach is 2,721.39 million kilometres.

PRINCIPAL CHARACTERISTICS

Basic information about the Uranian system is summarized in Table 17. The Uranian year is 84.01 Earth years. The eccentricity of the orbit is 0.0461, and the inclination of the orbit to the Earth's orbital plane is 0.774°. Low eccentricity and low inclination such as those of Uranus are characteristic of all planetary orbits. It is believed that collisions and gaseous drag took energy out of the orbits while the planets were forming and so reduced the eccentricities and inclinations to their present values. Thus, Uranus formed with the rest of the planets soon after the birth of the Sun.

The mass of Uranus, 8.684×10^{25} kilograms, is 85 percent that of Neptune. At the level where the atmospheric pressure is one bar (roughly equivalent to the Earth's sea-level pressure), the equatorial radius is 25,559 kilometres, which is 3.2 percent greater than that of Neptune. Thus Uranus and Neptune are nearly twin planets, although the difference in their respective bulk densities, 1.285 as compared with 1.64 grams per cubic centimetre, reveals a fundamental difference in composition and internal structure. Although Uranus and Neptune are significantly larger than the terrestrial planets, their radii are less than half those of the largest planets, Jupiter and Saturn. The equatorial surface gravity of Uranus is 8.69 metres per square second, which is 11 percent smaller than that of the Earth.

According to international convention, the north pole of a planet is defined as the pole that is above the equatorial plane of the solar system, regardless of the direction in which the planet is spinning. In terms of this definition, Uranus spins clockwise about its north pole, which is opposite to the spin of the Earth. When the Voyager 2 spacecraft flew by Uranus on Jan. 24, 1986, the north pole was in darkness and the Sun was almost directly overhead at the south pole. In 42 years (one-half the Uranian year), the Sun moves from being overhead at one pole to being overhead at the other. The 98° tilt is thought to have arisen during the final stages of planetary accretion when bodies comparable in size to the present planets collided in a series of violent events that knocked Uranus on its side.

Rotational periods

The rotational period of Uranus, 17.24 hours, was inferred when the Voyager spacecraft detected periodic radio emissions that were coming from charged particles trapped in the magnetic field. Direct measurements of the field showed that it is tipped at an angle of 58.6° relative to the rotation axis and that it rotates with the same period. Since the field is thought to be generated in the electrically conducting interior of the planet, the 17.24-hour period is assumed to be that of the interior. The rotation causes an oblateness, or flattening of the planet's poles such that the polar radius is 2.29 percent smaller than the equatorial radius. Winds in the atmosphere cause cloud markings to rotate at periods ranging from 18 hours to slightly more than 14 hours.

THE ATMOSPHERE

Molecular hydrogen (H_2) and helium (He) are the two main constituents of the atmosphere. Hydrogen is detectable from the Earth in the spectrum of sunlight scattered by the clouds of Uranus. The helium-to-hydrogen ratio was determined from the bending of the Voyager radio signal as it passed through the atmosphere. The ratio of the number of helium molecules to the total number of diatomic hydrogen and helium molecules is 0.15, while the corresponding mass ratio of helium to the total is 0.26. These values are consistent with the values inferred for the Sun and are significantly greater than the values inferred for the atmospheres of Jupiter and Saturn. It is assumed that all four of the giant planets have the same proportions of hydrogen and helium as the Sun but that the helium has settled toward the centres of Jupiter and Saturn. The processes that cause this settling have been shown not to operate on less massive planets like Uranus and Neptune.

Jet Propulsion Laboratory/National Aeronautics and Space Administration

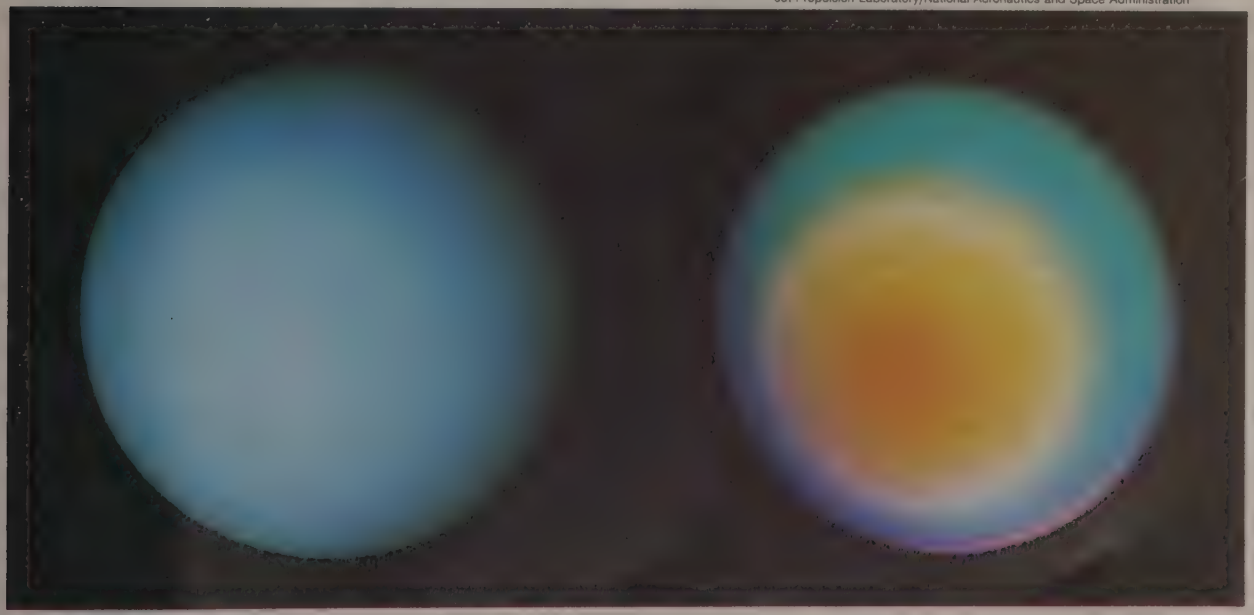


Figure 50: *Views of Uranus obtained by Voyager 2.*
(Left) The colour of this photograph was adjusted to simulate the view the human eye would normally see. (Right) The colour and brightness have been greatly exaggerated to enhance atmospheric features. The faint concentric bands are centred on the pole of rotation. (The small doughnut-shaped features are artifacts arising from dust in the camera.)

Table 17: Physical Data for Uranus

Mean distance from the Sun	2.87099×10^9 km (19.19 AU)
Orbital period (the Uranian year)	30,685 days (84.01 years)
Mean orbital velocity	6.81 km/s
Eccentricity of orbit	0.0461
Inclination of orbit to ecliptic	0.774°
Mass	8.684×10^{25} kg
Equatorial radius at one bar	25,559 km
Ellipticity	0.0229
Mean density	1.285 g/cm^3
Equatorial surface gravity	8.69 m/s^2
Equatorial escape velocity	21.3 km/s
Internal period of rotation	17.24 hours
Mean synodic period	369.66 Earth days
Inclination of equator to orbit	97.86°
Effective radiating temperature	59.1 K
Albedo	0.30
Magnetic dipole tilt relative to rotation axis	58.6°
Magnetic dipole strength at equator	0.228 gauss
Number of known satellites	15
Number of known rings	10

Methane (CH_4) absorbs strongly at near-infrared wavelengths, and it dominates that part of the spectrum even though the number of methane molecules is only 2.3 percent of the total. This estimate of the abundance was determined using Voyager 2 radio signals that probed to depths where the methane-to-hydrogen ratio is likely to be constant. If this constant is characteristic of the planet as a whole, the carbon-to-hydrogen (C:H) ratio of Uranus is 24 times that of the Sun. The large value of the C:H ratio suggests that the elements oxygen, nitrogen, and sulfur also are enriched relative to solar values. However, these elements are tied up in molecules of water (H_2O), ammonia (NH_3), and hydrogen sulfide (H_2S), which condense at levels below the part of the atmosphere that can be seen. Earth-based radio observations reveal a curious depletion of ammonia in the atmosphere, perhaps because hydrogen sulfide is more abundant and sequesters all the ammonia in cloud particles of ammonium hydrosulfide (NH_4SH). The Voyager ultraviolet spectrometer experiment detected traces of acetylene (C_2H_2) and ethane (C_2H_6) in abundances on the order of 10^{-8}. These gases are by-products of methane, which dissociates when ultraviolet light from the Sun strikes the upper atmosphere.

The average power radiated by Uranus is equivalent to that of a blackbody radiating at a temperature of 59.1 K. This radiation temperature is equal to the physical temperature of the atmospheric gases at a pressure of about 0.4 bar. Temperature decreases with height throughout this portion of the atmosphere up to the 70-millibar level, where it is about 52 K. The temperature rises from this point until it reaches 750 K in the exosphere (the top of the atmosphere at a distance of 1.1 Uranian radii from the planetary centre), where pressures are on the order of 10^{-12} bar. The cause of the high exospheric temperatures is uncertain, but it may involve a combination of ultraviolet absorption, electron bombardment, and inability of the gas to radiate at infrared wavelengths.

Atmospheric temperatures

Horizontal contrasts of temperature were measured by Voyager in two broad altitude ranges, one located in the range of 60 to 200 millibars and the other of 500 to 1,000 millibars. In both ranges the pole-to-pole temperature contrast is small, less than 1 K, despite the fact that the south pole was facing the Sun at the time of the Voyager encounter. This lack of global contrast is related to the efficient horizontal heat transfer and the large heat storage capacity of the deep atmosphere.

Although Uranus is nearly featureless, extreme contrast enhancement of the Voyager images reveals faint bands oriented parallel to circles of constant latitude (see Figure 50). Apparently the rotation of the planet and not the distribution of absorbed sunlight controls the cloud patterns. Rotation manifests itself through the Coriolis force, an effect that causes material moving on a rotating planet to appear to be deflected to either the right or the left depending on the sign of the latitude. Uranus therefore looks like a tipped-over version of Jupiter and Saturn in terms of cloud patterns.

The wind is the motion of the atmosphere relative to the rotating planet. At high latitudes on Uranus, as on the Earth, this relative motion is in the direction of the planet's rotation. At low (that is, equatorial) latitudes, the relative motion is in the opposite direction. On the Earth these directions are called east and west, respectively, but the more general terms are prograde and retrograde. The winds that exist on Uranus are several times stronger than those of the Earth. The wind is 200 metres per second (prograde) at a latitude of −55° and 110 metres per second (retrograde) at the equator. Neptune's equatorial winds are also retrograde, although those of Jupiter and Saturn are prograde. No satisfactory theory exists to explain these differences.

Winds

Uranus has no large spots like the Great Red Spot of Jupiter or the Great Dark Spot of Neptune. The measurements of the wind profile come from just four small spots whose contrast is no more than 2 or 3 percent relative to the surrounding atmosphere. Since the giant planets have no solid surfaces, the spots represent atmospheric storms. For reasons that are not clear, Uranus seems to have the smallest number of storms of any of the giant planets.

INTERIOR STRUCTURE AND COMPOSITION

Although Uranus has a somewhat lower density than Jupiter, it has a higher proportion of elements heavier than hydrogen and helium. Jupiter's greater mass (by a factor of 22) leads to a greater gravitational force and greater self-compression than on Uranus. If the two planets were made of the same material, Jupiter would be considerably denser than Uranus. Models proposed for the Uranian interior assume different ratios of rock (metals and their oxides), ice (water, methane, and ammonia), and gas (essentially hydrogen and helium). At the high temperatures and pressures within the giant planets, the "ices" will in fact be liquids. To be consistent with the bulk density data, the mass of rock plus ice must constitute roughly 80 percent of the total mass of Uranus, compared with 10 percent for Jupiter and 2 percent for a mixture of solar composition. In all models Uranus is a fluid planet, with the gaseous atmosphere gradually merging with the liquid interior. Pressure at the centre is about five megabars.

More information about the interior is obtained by comparing the model's response to centrifugal forces (as would arise from the planet's rotation) with the actual response measured as spacecraft are accelerated by the planet. This response is expressed in terms of the planet's oblateness. By measuring the degree of flattening at the poles as compared with the speed of rotation, one can infer the density distribution inside the planet. If two planets had the same mass and bulk density, the planet with most of its mass concentrated close to the centre would be less flattened by rotation. Before the Voyager mission, it was difficult to choose between models in which the three components rock, ice, and gas were separated into distinct layers and those in which the ice and gas were well mixed. From the large oblateness of Uranus combined with the planet's slow rotation, it now appears that the ice and gas are well mixed and a rocky core is small or nonexistent. The fact that the mixed model of Uranus fits the data better than the layered model may reveal information on the planet's formation. Instead of Uranus forming from a rock-ice core that subsequently captured gas from the solar nebula, the model seems to favour a process in which large, solid objects were continually captured into a giant planet that already contained major amounts of the gaseous component.

Uranus is different from the other giant planets in that it is not radiating a substantial amount of excess internal heat. The total heat output is determined from the measured infrared emissions, while the heat input is determined from the fraction of incident sunlight that is absorbed—*i.e.*, not scattered back into space. The fraction that is scattered is called the albedo, which for Uranus is about 0.3. The ratio of total power radiated to total power absorbed is between 1.00 and 1.06 for Uranus. (The equivalent ratios for the other giant planets are greater than 1.7.) Thus the internal heat source on Uranus is no more than 6 percent of the power absorbed. The small terrestrial planets generate relatively little internal heat; the comparable number for the Earth is only 10^{-4}.

Low internal heat output

It is not clear why Uranus has such a low internal heat output compared to the other Jovian planets. All the planets should have started warm, when gravitational energy was transformed into heat during planetary accretion. Over the age of the solar system, the Earth and the other smaller objects have lost most of their heat of formation. Being massive objects with cold surfaces, however, the giant planets store heat well and radiate poorly. Therefore, they should have retained large fractions of their heat of formation, which should still be escaping today. Chance events (collisions between large bodies) at the time of formation and the resulting differences in internal structure are one of the explanations that have been proposed to explain differences among the giant planets such as the anomalous heat output of Uranus.

MAGNETIC FIELD AND MAGNETOSPHERE

Like the other giant planets, Uranus has a magnetic field that is generated by convection currents in the electrically conducting interior. The magnetic dipole field, which resembles the field of a small but intense bar magnet, has a strength of 0.23 gauss in its equatorial plane at a distance of one Uranian equatorial radius (1 R_U) from the centre. The dipole axis is tilted with respect to the planet's rotation axis at an angle of 58.6°. This angle greatly exceeds that for the Earth (11°), Jupiter (9.6°), and Saturn (less than 1°). The magnetic centre is displaced by 0.3 R_U from the planet's centre. The displacement is mainly along the rotation axis toward the north pole.

The field is unusual not only because the dipole is tilted and offset from the centre of the planet but also because the small-scale components of the field are relatively large. This "roughness" suggests that the field is generated at shallow depths within the planet, because small-scale components of a field die out rapidly above the electrically conducting region. Thus the interior of Uranus must become electrically conducting closer to the surface than on Jupiter, Saturn, and the Earth. This inference is consistent with what is known about the internal composition, which must be mostly water, methane, and ammonia in order to match the average density of the planet. Since water and ammonia dissociate into positive and negative ions at relatively low pressures and temperatures, the field is expected to be generated close to the surface where these moderate conditions obtain.

Near-surface generation of the magnetic field

As with the other magnetic planets, the field repels the stream of charged particles (the solar wind) flowing outward from the Sun. The planetary magnetosphere—a huge cavity containing charged particles that are bound to the magnetic field—surrounds the planet and extends downstream from it. On the upstream side, the boundary between the magnetosphere and the solar wind is 18 R_U from the centre of the planet.

Because the largest Uranian satellites orbit within the magnetosphere, they absorb some of the trapped particles. The particles behave as if they were attached to the magnetic field lines, so that those lines intersecting a satellite in its orbit have fewer trapped particles than neighbouring field lines. The Uranian magnetosphere is composed of protons and electrons, indicating that the planet's upper atmosphere is supplying most of the material. There is no evidence of helium, which might originate from the solar wind, and no evidence of heavier ions, which might come from the Uranian satellites or their atmospheres.

Charged particles from the magnetosphere impinge on the upper atmosphere and stimulate the aurora. Auroral heating can just barely account for the high temperature (750 K) of the Uranian exosphere. One effect of the high temperature is that the atmosphere extends outward into the region occupied by ring particles and severely limits their orbital lifetime to values less than the age of the solar system.

THE SATELLITES AND RINGS

Uranus has 15 satellites and 10 narrow rings, all in nearly circular orbits with low inclinations relative to the equatorial plane of the planet. In general, the rings orbit closest to the planet, the smaller satellites orbit just outside the rings, and the larger satellites orbit farthest from the planet.

Satellites. Properties of the five major satellites are summarized in Table 18, and characteristics of the rings are given in Table 19. The 10 minor satellites, all of which were discovered by Voyager 2, have radii from 13 to 77 kilometres and orbit the planet at distances ranging from 49,750 to 86,000 kilometres. The innermost satellite, Cordelia, orbits just inside the outermost rings, Lambda and Epsilon.

Table 18: Major Satellites of Uranus

name	distance from centre of Uranus (km)	orbital period (days)	mean radius (km)	mean density (g/cm^3)
Miranda	129,780	1.414	236	1.15
Ariel	191,240	2.520	579	1.56
Umbriel	265,970	4.144	585	1.52
Titania	435,840	8.706	789	1.70
Oberon	582,600	13.463	761	1.64

The four largest satellites, Ariel, Umbriel, Titania, and Oberon, have densities that range from 1.5 to 1.7 grams per cubic centimetre, which is slightly greater than the density of a hypothetical satellite that is 60 percent ice and 40 percent rock. This hypothetical satellite would be obtained by cooling a mixture of solar composition and removing all the gaseous components. In contrast, Miranda and the smaller satellites of Saturn have densities that are slightly below the solar composition value, indicating a higher ice-to-rock ratio for these satellites.

Water ice shows up in the spectra of the five major satellites. At wavelengths outside the water bands, the spectral lines are neutral, indicating a gray colour. The Bond albedo, which is the ratio of sunlight reflected in all directions to incident sunlight, is low, ranging from 0.10 to 0.22 for the 5 large satellites and less than 0.1 for the 10 minor satellites. The obvious implication is that the surfaces consist of dirty water ice. The composition of the dark component is unknown. One possibility is carbon, originating from the inside of the planet or from the rings, which could release methane gas that later decomposes to produce solid carbon when bombarded by charged particles and solar ultraviolet light.

Low albedo

Two further observations indicate that the surfaces are porous and highly insulating. First, the reflectivity increases dramatically at opposition, when the observer is within 2° of the Sun as viewed from the planet. Such so-called opposition surges are characteristic of loosely stacked particles that shadow each other except in this special geometry. Second, the temperatures seem to follow the Sun during the day with no appreciable lag due to thermal inertia. Again, such behaviour is characteristic of porous surfaces.

Oberon and Umbriel display a dense population of large impact craters, similar to the lunar highlands and many of the oldest terrains in the solar system. In contrast, Titania and Ariel have far fewer craters in the large size range (50- to 100-kilometre diameter) but have comparable numbers in the smaller size ranges. The former craters are thought to date back more than four billion years to the early history of the solar system, while the latter are thought to reflect more recent events including, perhaps, secondary objects knocked loose from other satellites in the Uranian system. Thus the surfaces of Titania and Ariel are younger than those of Oberon and Umbriel. These differences do not follow an obvious pattern with respect to either distance from Uranus or satellite radius and are largely unexplained.

Volcanic deposits are generally flat, with lobate edges and surface ripples characteristic of fluid flow. Some of the deposits are bright, while some are dark. Since temperatures in the outer solar system are likely to be low, the erupting fluid was probably a water-ammonia mixture with a melting point well below the melting point of pure water ice. Brightness differences could reflect differences in the composition of the erupting fluid or differences in the history of the surface.

Riftlike canyons on all the Uranian satellites imply extension and fracturing of the surface. Miranda is the most

Figure 51: *Mosaic of Voyager 2 images of Miranda.*
The terrain of Miranda is characterized by extensive bands, ridges, and scarps. The unusual surface may be the result of a catastrophic impact that shattered the satellite followed by its reaccretion, in which subsurface materials became exposed and material formerly at the surface was buried in the interior. This view is centred on Miranda's south pole.
Jet Propulsion Laboratory/National Aeronautics and Space Administration

spectacular; some canyons are as much as 80 kilometres wide and 15 kilometres deep. The rupturing of the crust was caused by an expansion in the volume of the satellites, inferred to be in the range of 1 to 2 percent, except for Miranda, for which the expansions are 6 percent. Miranda's expansion could be explained if all the water in the satellite were once liquid and then froze in its interior after the crust had formed. Freezing under low pressure, the water would have expanded and thereby stretched the surface. The presence of liquid water at any stage of the satellite's history seems unlikely, however.

Unique appearance of Miranda

Miranda (Figure 51) has the appearance of an object that formed from separate pieces that did not totally merge. The basic surface is heavily cratered, but it is interrupted by three lightly cratered regions called coronae. These are fairly squarish, roughly one satellite radius on a side, and are surrounded by parallel bands that curve around the edges. The boundaries where the coronae meet the cratered terrain are sharp. The coronae are unlike any features found elsewhere in the solar system. Whether they reflect the satellite's heterogeneous origins, a giant impact that shattered the satellite, or a new pattern of eruption from the interior is not known.

Rings. The rings of Uranus were discovered from the Earth during stellar occultations—*i.e.*, when the planet passed between a star and the Earth, thereby momentarily blocking the star's light. The rings are narrow and fairly opaque (Table 19). Observed widths are simply the radial distances between the beginning and end of the occultation. Equivalent widths are the product (more precisely, the integral) of the radial distance and the fraction of starlight removed. The fact that the equivalent widths are generally less than the observed widths indicates that the rings are not completely opaque.

Combining the brightness in Voyager images with the equivalent widths from occultations, one finds that the ring particles reflect only 1 or 2 percent of the incident sunlight. The reflectance spectrum is nearly flat. Ordinary soot, which is mostly carbon, is the closest terrestrial analogue. It is not known whether the carbon comes from darkening of methane by particle bombardment or is intrinsic to the ring particles.

Propagation of the Voyager radio signal through the rings to the Earth reveals that the rings consist of mostly large particles with radii greater than 70 centimetres. Scattering of sunlight at large phase angles, when the rings are almost between the Sun and the Voyager camera, also reveals small dust particles, whose size is comparable to the wavelength of light. Only a small amount of dust is found in the main rings. Most microscopic-size particles are instead distributed in the spaces between the main rings, suggesting that the rings are losing mass as a result of collisions. The lifetime of the dust in orbit around Uranus is limited by drag exerted by the planet's extended atmosphere and by light pressure; the dust particles are driven to lower orbits and eventually fall into the Uranian atmosphere. The calculated orbital lifetimes are so short (1,000 years) that the dust must be rapidly and continually created.

Collisions between the tightly packed ring particles would naturally lead to an increase in the radial width of the

Table 19: Rings of Uranus

name	distance from centre of Uranus (km)	equivalent width* (km)	observed widths† (km)
6	41,837	0.66	1–2
5	42,235	1.23	2–7
4	42,571	1.06	1–6
Alpha	44,718	3.86	4–11
Beta	45,661	3.16	4–13
Eta	47,176	0.64	1–4
Gamma	47,627	3.13	2–8
Delta	48,300	2.69	3–8
Lambda	50,023	0.1	2–3
Epsilon	51,149	42.8	20–95

*Equivalent width is the product of width times fraction of light attenuated and is given for visible light. †The range of observed widths reflect real variations with respect to longitude as well as measurement error.

Figure 52: *Shepherd satellites of the Epsilon ring.* The two satellites Ophelia (1986U8) and Cordelia (1986U7) orbit outside and inside, respectively, of the Epsilon ring, thereby confining the ring particles through resonant gravitational interactions.

Jet Propulsion Laboratory/National Aeronautics and Space Administration

Shepherd satellites

rings. Satellites more massive than the rings can halt this spreading in a process known as shepherding. At certain radii, termed resonances, in which the satellite's orbital period is a whole number ratio of the ring particles' orbital period, the satellite exerts a net torque on the ring by gravitational interaction. As the satellite and ring exchange angular momentum, energy is dissipated by collisions among the ring particles. The outcome of this interaction is that the satellite and ring repel each other. The one in the outer orbit moves outward, and the one in the inner orbit moves inward. Since the satellite is much more massive than the ring, it prevents the ring from spreading across the radius at which resonance occurs. A pair of satellites on either side of a ring can maintain its narrow width.

The inner two satellites, Cordelia and Ophelia, orbit on either side of the Epsilon ring at exactly the right radii required for shepherding (see Figure 52). Shepherds for the other rings have not been found, perhaps because the shepherds are too small to be seen in the Voyager images. Small moons may also be the reservoir that supplies the dust that leaves the ring system. Atmospheric drag is so large that the rings may be short-lived, in which case their origin and history are still unknown.

HISTORY OF OBSERVATION

Uranus was discovered in 1781 by the English astronomer William Herschel, who had undertaken a survey of all stars down to eighth magnitude—*i.e.,* those about five times fainter than stars visible to the naked eye. On March 13, 1781, he found "a curious either nebulous star or perhaps a comet," distinguished from the stars by its clearly visible disk. Its lack of any trace of a tail and its slow motion led to the suggestion that the observations were consistent with a planet, rather than a comet or asteroid, moving in a nearly circular orbit. Observations of the orbit during the next 65 years revealed discrepancies—gravitational forces on Uranus that were not due to any known planet and which ultimately led to the discovery of Neptune in 1846.

Herschel suggested naming the new planet the Georgian Planet after his patron, King George III of England, while the French favoured the name Herschel. The planet was eventually named according to the tradition of naming planets for the gods of Greek and Roman mythology; Uranus is the father of Saturn, who is in turn the father of Jupiter.

The orbit of Uranus seemed to fit a simple equation, Bode's law, that had been developed in 1766 and was popularized in 1772 to explain the orbits of the six planets known to the ancients. In addition, asteroids seemed to fill a gap where the law predicted another planet. For about three-quarters of a century, these successes overcame doubts stemming from the fact that the law had no theoretical basis and provided only an approximate fit to the planetary orbits. Neptune did not fit the pattern at all (being 21 percent closer to the Sun than predicted), nor did Pluto, and now Bode's law is of only historical significance.

After the discovery in 1781, Herschel continued to observe the planet with larger and better telescopes and eventually discovered the outer two satellites, Titania and Oberon, in 1787. Two more satellites, Ariel and Umbriel, were discovered by the British astronomer William Lassell in 1851. The names of the four satellites come from English literature, three taken from Shakespeare, and were proposed by Herschel's son, John Herschel. A fifth satellite, Miranda, was discovered by Gerard P. Kuiper in 1948. The tradition of naming the satellites after characters in Shakespeare's plays continues to the present.

Discovery of major satellites

Since the 1860s the spectrum of Uranus has been known to contain deep absorption bands in the red region of the electromagnetic spectrum. In 1932 the German-American astronomer Rupert Wildt showed that these must be due to methane in the planet's atmosphere. The Canadian physicist Gerhard Herzberg first detected hydrogen in the atmospheres of both Uranus and Neptune in 1952. The rings were discovered by James L. Elliot of the United States in 1977 while he was observing the planet pass in front of a bright star. They showed up as places where the starlight briefly faded at some considerable distance above the atmosphere. The somewhat cumbersome nomenclature (Table 19) arose as subsequent observations

By courtesy of the Jet Propulsion Laboratory/National Aeronautics and Space Administration

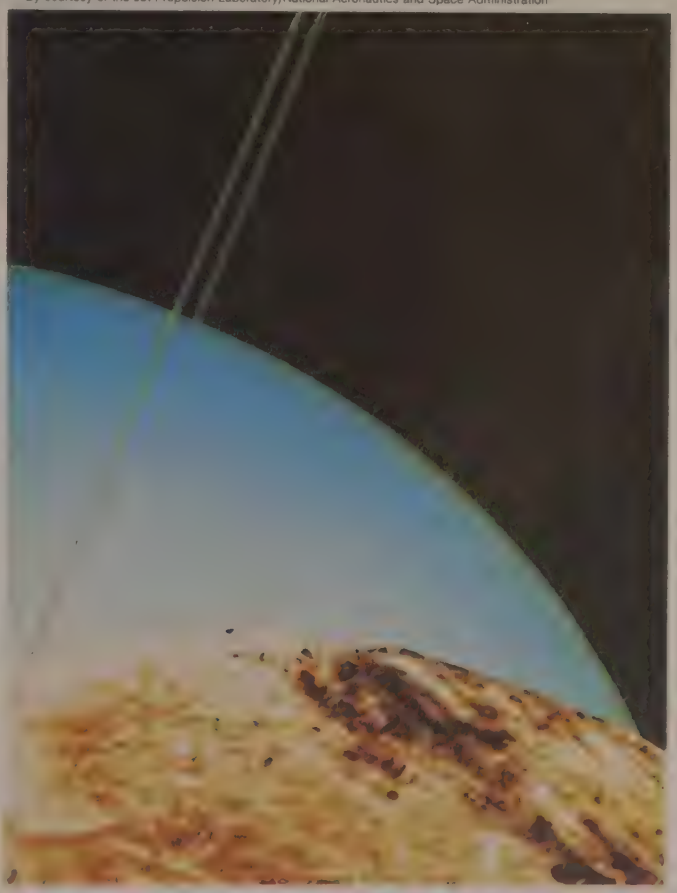


Figure 53: Montage of Voyager 2 photographs taken in January 1986 that simulates a view of Uranus and rings as if seen over the horizon of Miranda, one of the satellites of Uranus.

revealed new rings in places that did not fit the original nomenclature.

The Voyager 2 spacecraft flew by Uranus on Jan. 24, 1986. It was the first man-made object to visit the planet, and it revealed much about the Uranian system. A montage of Voyager 2 photographs (Figure 53) shows Uranus and its rings in a simulated view as if seen from the horizon of Miranda. Voyager provided an accurate determination of the masses and radii of the planet and its major satellites. It detected the magnetic field and determined its strength and orientation. Voyager instruments measured the rotation rate of the field and hence of the planet itself. Images from Voyager revealed for the first time the weather patterns in the atmosphere and the surface morphology of the satellites. Voyager discovered 10 new satellites, a narrow ring, and dust bands between the narrow rings. It provided details of ring structure at scales unachievable from the Earth. Yet, despite these advances, many unanswered questions remain that can be answered only by another spacecraft or by a major advance in Earth-based observations. (A.P.I.)

Neptune

Neptune, the eighth planet in average distance from the Sun, was named for the Roman god of the sea. The sea-god's three-pronged trident (Ψ) serves as its astronomical symbol. Neptune's distance from the Sun varies between 29.8 and 30.4 astronomical units (AUs). Its diameter is nearly four times that of the Earth, but because of its great distance Neptune cannot be seen from the Earth without the aid of a telescope. Neptune's deep blue colour is due to the absorption of red light by methane gas in its atmosphere. It receives less than half as much sunlight as Uranus, but heat escaping from its interior makes Neptune slightly warmer than the latter. The heat liberated may also be responsible for Neptune's stormier atmosphere, which exhibits the fastest winds seen on any planet in the solar system.

PRINCIPAL CHARACTERISTICS

Neptune's orbital period is 164.8 Earth years. It has not completely circled the Sun since its discovery in 1846, so some refinements in its orbital size and shape are still expected. Voyager 2's encounter with Neptune in 1989 resulted in an upward revision of about 0.17 percent in its estimated mean distance from the Sun, which is now thought to be 4,504,300,000 kilometres. Its orbital eccentricity of 0.009 means that Neptune's orbit is very nearly circular; among the nine planets in the solar system, only Venus has a smaller eccentricity. Neptune's seasons (and the seasons of its moons) are therefore of nearly equal length, each more than 41 Earth years in duration. The tilt of Neptune's equator relative to its orbit is 29.6°, somewhat larger than the Earth's 23.4°.

41-year seasons

The length of Neptune's day, as determined by Voyager 2, is 16.11 hours. Its equatorial diameter, measured at the one-bar pressure level (the pressure of the Earth's atmosphere at sea level), is 49,528 kilometres. Because of polar flattening, Neptune's polar diameter is only 48,680 kilometres. Neptune's volume is equivalent to 57.7 Earth volumes; Uranus is slightly larger with a value of 63 Earth volumes. Owing to its greater density (1.64 grams per cublic centimetre), however, Neptune's mass is 18 percent higher than the mass of Uranus. Neptune has a mass equivalent to that of 17.15 Earth masses.

Table 20: Physical Data for Neptune

Mean distance from the Sun	4,504,300,000 km (30.058 AU)
Eccentricity of orbit	0.009
Inclination of orbit to ecliptic	1.77°
Sidereal period of revolution	164.8 Earth years
Rotational period	16.11 hours
Radiometric bond albedo	0.29 ± 0.07
Effective temperature	59.3 K
Energy balance (emitted/absorbed)	2.7
Mean synodic period	367.49 days
Mean orbital velocity	5.43 km/s
Mean daily motion	0.005981°
Inclination of equator to orbit	29.6°
Mass	17.15 Earth masses
Equatorial radius at one bar	24,764 km
Polar radius at one bar	24,340 km
Mean density	1.64 g/cm³
Apparent magnitude	+7.8
Number of known satellites	8
Number of known rings	4
Magnetic dipole moment	0.133 gauss R^3_N
Magnetic dipole tilt	46.8
Magnetic dipole offset	0.55 R_N

THE ATMOSPHERE

As with the other giant planets of the outer solar system, Neptune's atmosphere is composed predominantly of hydrogen and helium. Near the one-bar pressure level in the atmosphere, these two gases contribute nearly 98 percent of the atmospheric molecules. Most of the remaining molecules consist of methane gas. Hydrogen and helium are nearly invisible, but methane strongly absorbs red light. Sunlight reflected off Neptune's clouds therefore exits the atmosphere with most of its red colours removed: this effect is responsible for Neptune's blue appearance (see Figure 54).

The temperature of Neptune's atmosphere varies with altitude. A minimum temperature of about 50 K occurs at pressure near 0.1 bar. The temperature increases with altitude to about 750 K at 2,000 kilometres (corresponding to a pressure of 10^{-11} bar) and remains uniform above that altitude. Temperatures increase with depth below the 0.1-bar level to about 7,000 K near the centre of the planet, where the pressure may reach 5,000,000 bars. The effective temperature of Neptune (the temperature of a perfect blackbody emitter of the same cross section) is 59.3 K.

Neptune is more than 50 percent farther from the Sun than is Uranus. Neptune consequently receives less than half as much sunlight as the latter. Yet the effective temperatures of these two giant gaseous planets are nearly equal. Uranus and Neptune each reflect (and hence also absorb) about the same proportion of the sunlight that reaches them. For reasons not fully understood, Neptune emits more than twice as much energy as it receives from the Sun. The added energy is generated in Neptune's interior. Uranus, by contrast, has little energy escaping from its interior.

At the reference level of one bar, the mean temperature of Neptune's atmosphere is roughly 74 K. Atmospheric temperatures are a few degrees warmer at the equator and poles than at mid-latitudes. This is probably an indication that air currents are rising near mid-latitudes and descending near the equator and poles. This vertical flow may extend to great heights within the atmosphere. A more vertically confined horizontal wind system exists near the cloud tops. As with the other giant planets of the outer solar system, the winds are constrained to blow generally along lines of constant latitude and are relatively invariable with time. Winds on Neptune vary from about 100 metres per second (m/s) in an easterly (prograde) direction near latitude 70° S to as high as 700 m/s in a westerly (retrograde) direction near latitude 20° S. These 700-m/s atmospheric winds are the highest measured anywhere in the solar system.

High atmospheric winds

The high winds and relatively large contribution of escaping internal heat may be responsible for the observed turbulence in Neptune's visible atmosphere. Two large dark ovals are clearly visible in images of Neptune's southern hemisphere (see Figure 54). The largest, called the "Great Dark Spot" because of its similarity in latitude and shape to Jupiter's Great Red Spot, is comparable to the entire Earth in size. It is near this Great Dark Spot that the highest wind speeds were measured. A somewhat smaller "Small Dark Spot" circles the planet near latitude 55° S. These two atmospheric storms may be centres where strong upwelling of gases from the interior takes place. A bright feature, dubbed the "Scooter," consists of a series of small streaks that vary in number and size over time, causing the Scooter to change in shape. Several other bright features, such as the bright white companions of the two dark spots, may be attributable to methane ice clouds created by strong upward motions of pockets of methane gas to higher, colder altitudes in the atmosphere.

Neptune is the only one of the giant planets of the solar

Figure 54: *Voyager 2 image of clouds in Neptune's atmosphere.*
The Great Dark Spot and its associated bright methane-ice clouds are shown slightly to the left of centre, the small dark spot is at the lower left, and the bright "Scooter" is just northward of the small dark spot.
Jet Propulsion Laboratory/National Aeronautics and Space Administration

system to display cloud shadows cast by high dispersed clouds on a lower, more continuous cloud bank. The higher clouds, probably composed of methane ice crystals, are generally located from 50 to 100 kilometres above the main cloud deck, which may be composed of ammonia or hydrogen sulfide ice crystals. As with the other gas giants, deeper cloud layers, invisible to the remote sensing instruments carried by the Voyager spacecraft, are thought to exist, but their composition is dependent on the relative amounts of gases composed of compounds of sulfur and nitrogen. Clouds of water ice are expected to occur at depths within Neptune's atmosphere where the pressure is in excess of 100 bars.

INTERIOR STRUCTURE AND COMPOSITION

Neptune's mean density is slightly less than 30 percent that of the Earth; nevertheless, it is the densest of the giant planets. Neptune's greater density implies that a larger percentage of its interior is composed of melted ices and molten rocky materials than is the case for the other gas giants.

The distribution of these heavier elements and compounds is poorly known at present. Voyager 2 data suggest, however, that the planet is unlikely to have a distinct inner core of molten rocky materials surrounded by an outer core of melted ices of methane, ammonia, and water. The relatively long rotational period of Neptune (16.11 hours) was about one hour longer than would be expected from such an interior model. Scientists have concluded that the heavier compounds and elements must not be centrally condensed; instead, they may be spread almost uniformly throughout the interior. In this respect, as in many others, Neptune resembles Uranus far more than the larger giants Jupiter and Saturn.

Distribution of rock and ice

The large fraction of Neptune's total heat budget derived from the planet's interior may not necessarily imply that Neptune is hotter at its centre than Uranus. Multiple stratified layers in the deep atmosphere of Uranus may serve to insulate the interior, trapping within Uranus the radiation that more readily escapes from Neptune.

MAGNETIC FIELD AND MAGNETOSPHERE

Neptune, like most of the planets in the solar system, possesses an internal magnetic field. The Earth's magnetic field is thought to be generated by electrical currents flowing in its liquid iron core, and electrical currents flowing within the outer cores of liquid metallic hydrogen in Jupiter and Saturn may similarly be the source of their magnetic fields. All three have magnetic fields relatively well centred within the planets and aligned within about 10° of their rotation axes. Uranus and Neptune, by contrast, have magnetic fields that are tilted with respect to their rotation axes by 59° and 47°, respectively. Furthermore, these magnetic fields are not well centred within their planets. Uranus' field is offset 30 percent of the distance from its centre to its cloud tops. Neptune's field, with an offset of 55 percent of Neptune's radius, originates in a portion of the interior that is actually closer to the cloud tops than to the centre. The unusual configurations of the magnetic fields of Uranus and Neptune have led scientists to speculate that these fields may be generated in processes occurring in the upper layers of the planetary interiors rather than near their centres.

Because of the high tilt of Neptune's magnetic field, charged particles (predominantly protons and electrons) trapped in the magnetosphere are repeatedly swept past the orbits of the satellites and rings. Many of these charged particles may be absorbed by the satellites and rings, effectively emptying from the magnetosphere a large fraction of its charged particle content. Neptune's magnetosphere is populated with fewer protons and electrons per unit volume than that of any of the other gas giant planets.

THE SATELLITES AND RINGS

Satellites. Prior to the Voyager 2 encounter in August 1989, Neptune's only known satellites were Triton and

Nereid. Triton is the lone large moon in the solar system to travel backward (in a direction opposite to the planet's rotation) around its primary. Among the largest satellites in the solar system, inclinations are less than about 5°; Triton's orbit, however, is tilted 157° to Neptune's equator. Nereid is very distant from the planet and has the most eccentric orbit of any known natural satellite. At its apoapsis (greatest distance), Nereid is nearly seven times as far from Neptune as at its periapsis (smallest distance). Even at its closest approach, Nereid is nearly four times the distance of Triton, Neptune's second most distant satellite.

Voyager 2 observations added six previously unknown satellites to Neptune's system. All are less than half Triton's distance from Neptune and travel in nearly circular orbits that are prograde and in or near Neptune's equatorial plane. The names of the satellites selected by the International Astronomical Union are all derived from Greek and Roman mythology and correspond to minor gods who served Neptune. These names, in order of increasing distance from Neptune, are Naiad, Thalassa, Despina, Galatea, Larissa, and Proteus. Physical data on the eight satellites are summarized in Table 21.

Table 21: Satellite Data

satellite name	radius (km)	density (g/cm³)	distance (km)*	inclination (°)	eccentricity	orbital period (days)
Naiad	29	—	48,230	4.74	0.0003	0.2944
Thalassa	40	—	50,070	0.21	0.0002	0.3115
Despina	74	—	52,530	0.07	0.0001	0.3347
Galatea	79	—	61,950	0.05	0.0001	0.4287
Larissa	96	—	73,550	0.20	0.0014	0.5547
Proteus	208	—	117,640	0.04	0.0004	1.1223
Triton	1,350	2.07	354,800	156.8	0.000	5.8768
Nereid	170	—	5,509,100	27.6	0.753	359.632

*From Neptune's centre.

Five of the six recently discovered satellites (all but Proteus) orbit Neptune in less time than the 16.11-hour rotational period of the planet. Hence, to an observer positioned near Neptune's cloud tops, these five would appear to rise in the west and set in the east. Only two of the six "new" satellites were seen from close enough range to detect both their size and approximate shape. Proteus and Larissa are irregular in shape and appear to have heavily cratered surfaces. The sizes of the other four are estimated from their integrated brightness by assuming that their surface reflectances are similar to those of Proteus and Larissa—namely, about 7 percent. Proteus, with a radius of approximately 208 kilometres, is slightly larger than Nereid (with a radius of about 170 kilometres). The other five new satellites are much smaller, each having an average radius of less than 100 kilometres.

Nereid was not observed from close range, but Voyager 2 data indicate a nearly spherical shape for the outermost of Neptune's satellites. Voyager detected no large variations in brightness as Nereid rotates. The highly elliptical orbit makes it unlikely that Nereid's rotational and orbital periods are equal, but Voyager 2 was not able to determine a rotational period. Rotational periods for all the other satellites (including Triton) are probably equal (or very nearly equal) to their orbital periods.

Pre-Voyager estimates of Triton's size made from the Earth were based on an erroneously high mass determination and assumption of low surface reflectivity. Triton's mass is now known to be only a small fraction of the previously accepted value, and the average surface reflectivity is high. Triton's radius as measured by Voyager 2 is 1,350 kilometres. Triton has a highly reflective icy surface, in contrast to the Moon's dark surface devoid of volatile components. Triton's low mass is likely a consequence of a predominantly water-ice interior surrounding a denser rocky core, while the Moon's composition is that of almost exclusively rocky materials. Nevertheless, Triton's mean density of 2.07 grams per cubic centimetre is higher than that measured for any of the satellites of Saturn or Uranus and is surpassed among large satellites only by the Moon and Jupiter's satellites Io and Europa.

Triton's visible surface is covered by methane and nitrogen ices. Evidence of trace amounts of carbon monoxide and carbon dioxide ices also has been revealed in spectroscopic studies conducted from the Earth. Even at the remarkably low 38 K surface temperature, sublimation of nitrogen ice is sufficient to form a tenuous atmosphere whose near-surface pressure is less than 0.00002 bar. A polar ice cap, presumably composed of nitrogen ice deposited in the prior winter, covered most of the southern hemisphere of Triton during Voyager 2's flyby in 1989. At that time Triton was nearly three-quarters of the way through its 41-year southern springtime. Equatorward of the polar cap, much of the terrain had the appearance of a cantaloupe rind, consisting of dimples crisscrossed with a network of fractures. This unique terrain is shown in Figure 55.

Within the polar cap region, numerous darker streaks provide evidence of surface winds. At least two of the streaks, and perhaps dozens, are the result of active, geyser-like plumes erupting during the Voyager 2 flyby. Nitrogen gas, escaping through vents in the overlying ice, carries entrained dust particles to heights of about 8 kilometres, where the dust is then transported downwind to distances of up to 150 kilometres. The energy sources and mechanisms for driving these plumes are not yet well understood, but their preference for latitudes illuminated vertically by the Sun has led to the conclusion that incident sunlight is an important factor in the process.

Triton's geyserlike plumes

Near the equator on the Neptune-facing side of Triton exist at least two and perhaps several frozen lakelike features with terraced edges. The terraced edges are probably the result of multiple epochs of melting, with each successive melt involving a somewhat smaller surface area of ice. The vertical extent of some of the cliffs (terrace edges) is more than one kilometre. Even at Triton's low surface temperature, nitrogen or methane ices are not strong enough to support a structure of that height without structural failure. It is assumed that the underlying material supporting these structures is water ice, although no direct evidence for water ice is seen in the spectra of Triton. A thin veneer of nitrogen or methane ice could effectively hide the spectral signature of water ice.

Triton is similar in size, density, and surface composition to the planet Pluto. It is thought to be a captured satellite, perhaps formed originally as an independent planet in the outer solar system. At some point in Neptune's early history, Triton's orbit may have carried it too near the gas giant. Gas drag in Neptune's extended atmosphere, or a collision with an existing satellite of Neptune, slowed Triton enough to place it in an elongated orbit, backward and tilted with respect to the orbits of the previously existing satellites. As Triton raised tides in Neptune's atmosphere, these tides in turn selectively retarded Triton in the closer portions of its orbit, eventually circularizing its path around Neptune. This process from capture to circular orbit may have taken more than one billion years, during which time the enormous tidal forces most likely melted the entire interior of Triton. The molten body would have undergone differentiation, with the denser material sinking into a core region and the more volatile materials rising to the surface.

It is thought that Triton's surface cooled faster than its interior and formed a thick layer of predominantly water ice. As subsurface water froze, it expanded, fracturing the outer ice layer and flowing through and filling the cracks. The intersecting fractures visible in Voyager images of Triton's surface provide strong corroborating evidence for the existence of water ice inside this satellite, since no other compositional candidates for Triton's subsurface expand as they freeze.

The process of capture and circularization of Triton's orbit would have severely disrupted any previously existing satellite system formed at the same time as Neptune. Nereid's radical orbit may be one consequence of this process. Satellites orbiting between Nereid and the six satellites discovered by Voyager 2 would have been ejected from the Neptune system, thrown into Neptune itself, or absorbed by the molten Triton. Those satellites orbiting at a distance of less than half Triton's present distance from

Figure 55: *Voyager 2 image of Triton.*
This 14-frame composite image shows the satellite's near-equatorial cantaloupe terrain and fracture network, the remnant southern hemisphere ice coverage, and the dark wind streaks within the ice-cap region.
Jet Propulsion Laboratory/National Aeronautics and Space Administration

Neptune would not have been close enough to the planet to avoid some disruption. Thus, the present orbits of the six small inner satellites are probably very different from their original orbits, and they may only be fragments of original satellites formed along with Neptune. Subsequent bombardment of these satellites by Neptune-orbiting debris and by meteorites may have further altered their sizes, shapes, and orbits.

Rings. Neptune's system of narrow rings (Figure 56), which may be a relatively new addition to the planet's family, displays a quite unusual feature. Although narrow rings have been seen around each of the other three gas giants, none have displayed the striking nonuniformity of particle density of Neptune's outermost ring, provisionally called 1989N1R. Material is clumped in at least five brighter regions, called arcs, that are found within a 45° segment of the ring. They range in length from about 1,000 kilometres to more than 10,000 kilometres. Although Galatea may gravitationally interact with 1989N1R to temporarily trap ring particles in such arclike regions, collisions between ring particles should eventually spread the constituent material relatively uniformly around the circumference of 1989N1R. Azimuthal uniformity is characteristic of almost all the known planetary rings, although some of the elliptical rings of Uranus exhibit smooth variations in width. 1989N1R's enigmatic arcs may be the

Ring arcs

Jet Propulsion Laboratory/National Aeronautics and Space Administration

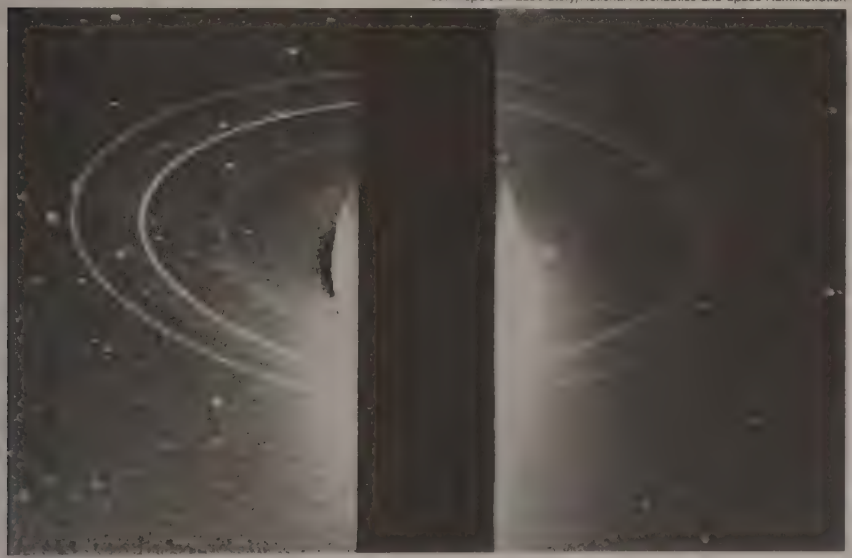


Figure 56: *Neptune's ring system.*
Two long-exposure images obtained by Voyager 2 a few hours after the spacecraft's closest approach to the planet show the main features of Neptune's rings. (The bright arcs of the outer ring are not visible here because they were on the opposite side of the planet when each photograph was taken.) The overexposed image of Neptune's southern polar crescent is near the centre.

result of the breakup of a small satellite within the past few thousand years.

The inner rings of Neptune (1989N2R, 1989N3R, and 1989N4R) lack the unusual azimuthal nonuniformity exhibited by 1989N1R. 1989N2R is narrow—less than 15 kilometres in radial width—and closely resembles the nonarc regions of 1989N1R. The small satellites Galatea and Despina orbit Neptune just planetward of 1989N1R and 1989N2R, respectively, and may gravitationally repel particles near the inner edges of these rings. 1989N3R is much broader and fainter than either 1989N1R or 1989N2R, possibly owing to the absence of any strong particle-shepherding action from a nearby satellite. 1989N4R consists of a faint plateau of ring material that extends outward in radial distance from 1989N2R about halfway to 1989N1R.

None of Neptune's rings were detected at radio wavelengths, indicating that they are nearly devoid of particles in the centimetre or larger size range. The fact that the rings are most visible when backlighted by sunlight implies that they are largely populated by dust-sized particles. Their chemical makeup is not known, but, like the rings of Uranus, the ring-particle surfaces (and possibly the particles in their entirety) may be composed of radiation-darkened methane ices. Physical data on the rings are summarized in Table 22.

Table 22: Ring Data

ring designation	common name	distance (km)*	comments
1989N3R	Galle	41,900	about 1,700 km wide with indistinct edges
1989N2R	Le Verrier	53,200	radial width not resolved; flanked by Despina
1989N4R	"Plateau"	56,100	about 5,800 km wide with inner edge at 1989N2R
1989N1R	Adams	62,900	15 km wide; bright arcs; flanked by Galatea

*From Neptune's centre.

HISTORY OF OBSERVATION

Brightness of Neptune

Neptune is the only giant gaseous planet that is not visible without a telescope. With an apparent brightness of +7.8 in stellar magnitudes, it is approximately one-fifth as bright as the faintest stars visible to the naked eye. Hence, it is relatively certain that there were no observations of Neptune prior to the use of telescopes. Galileo is credited as the first to view the heavens with a telescope in 1609. His sketches made a few years later suggest that he may have seen Neptune when it passed near Jupiter. He failed, however, to recognize it as a planet and therefore cannot be credited with its discovery.

Prior to the discovery of Uranus by William Herschel in 1781, the consensus among astronomers and philosophers alike was that all the planets in the solar system were those six that had been observed since ancient times. The discovery of a seventh planet almost immediately led astronomers and others to suspect the existence of still more planetary bodies. Additional impetus was given by a mathematical curiosity that has come to be known as Bode's law or the Titius-Bode law. In 1766 Johann Daniel Titius of Germany noted that the then-known planets formed an orderly progression in distance from the Sun that could be expressed as a simple mathematical equation. In astronomical units (AU), Mercury's distance is very nearly 0.4; the distances of Venus, the Earth, Mars, Jupiter, and Saturn are approximately $0.4 + (0.3 \times 2^n)$, where n is 0, 1, 2, 4, and 5, respectively, for the five planets. The astronomer Johann Elert Bode, also of Germany, published the law in 1772 in a popular introductory astronomy book, proposing that the missing 3 in the progression might indicate that a planet between Mars and Jupiter remained to be discovered.

The suggestion was met with little enthusiasm until it was discovered that Uranus was at a distance (19.2 AU) very nearly equal to that predicted by Bode's law (19.6 AU) for $n = 6$. It was found, however, that a group of asteroids orbiting between Mars and Jupiter, rather than a new planet, satisfied the $n = 3$ case of the equation (see below *Other constituents of the solar system: Asteroids: Historical survey of major asteroid discoveries*).

Some astronomers were so impressed by the seeming success of Bode's law that they proposed the name Ophion for a large planet predicted to lie beyond Uranus at a distance of 38.8 AU. In addition to the scientifically unfounded prediction from Bode's law, observations of Uranus provided more direct evidence for the existence of another planet. Uranus was not following the path predicted by the laws of motion and the gravitational forces exerted by the Sun and the known planets. Furthermore, more than 20 prediscovery sightings of Uranus dating back as far as 1690 also disagreed with the calculated positions of Uranus for the respective time at which each observation was made.

In 1843 the British mathematician John Couch Adams began a serious study to see if he could predict the location of a more distant planet that could account for the strange motions of Uranus. Adams succeeded and communicated the results of his study to the astronomer royal, Sir George B. Airy, but Airy did not take them seriously. In 1846 Urbain-Jean-Joseph Le Verrier of France, unaware of Adams' efforts in Britain, began a similar study of his own.

By mid-1846 the astronomer John Herschel, son of William Herschel, expressed his opinion that the mathematical studies under way could well lead to the discovery of a new planet. Airy, convinced by Herschel's arguments, sent Adams' calculations to James Challis at the Cambridge Observatory. Challis began a systematic search of a large area of sky surrounding Adams' predicted location. The search was slow and tedious, because Challis had no detailed maps of the dim stars in the area where the new planet was predicted. He would draw charts of the stars he observed and then compare them with the same region several nights later to see if any of the "stars" had moved.

Le Verrier also had difficulty convincing astronomers in his country that a telescopic search of the skies in the area he predicted for the new planet was not a waste of time. On Sept. 23, 1846, he communicated his results to the Berlin astronomer Johann Gottfried Galle. Galle and his assistant Heinrich Louis d'Arrest had access to detailed star maps of the sky painstakingly constructed to aid in the search for new asteroids. Galle and d'Arrest identified Neptune as a noncharted star that same night and verified the next night that it had moved relative to the background stars.

Neptune's identification in the sky

Although Galle and d'Arrest have the distinction of being the first individuals to identify Neptune in the night sky, credit for its "discovery" more properly belongs to Adams and Le Verrier for their accurate calculations of Neptune's position. At first the French attempted to proclaim Le Verrier as the sole discoverer of the new planet, suggesting that the planet be called Le Verrier in his honour. The suggestion was not favourably received outside France, both because of the studies by Adams and because of the general reluctance to name a major planet after a living individual. Neptune's discovery was eventually credited to both Adams and Le Verrier, and the more traditional practice of using names from ancient mythology for planets eventually prevailed.

It is interesting to note that Bode's law, which played a significant role in the search for additional planets, was finally laid to rest by the discovery of Neptune. Instead of being near the predicted 38.8 AU, Neptune was found to be only 30.1 AU from the Sun. This discrepancy, combined with the lack of any scientific foundation, discredited the law. The subsequent discovery of Pluto at a distance of 39.6 AU was even more at variance with the formula's prediction of 77.2 AU for $n = 8$. Not even the proximity of Pluto's mean distance to the 38.8 AU predicted for $n = 7$ could resurrect the credibility of Bode's law.

Pre-Voyager observations of Neptune often suffered greatly as a consequence of the planet's enormous distance from both the Earth and the Sun. Its distance of 30.1 AU from the Sun means that the brightness of sunlight incident upon its satellites and upon its upper atmosphere is barely 0.1 percent of that at the Earth. Telescopic viewing of Neptune from the bottom of the Earth's atmosphere cannot resolve features smaller than about one-tenth of Neptune's diameter, even under the best observ-

ing conditions. Most of the early telescopic observations concentrated on determining Neptune's size and orbital parameters and searching for satellites.

Following the successful detection of rings around Uranus in 1977, similar attempts were made to use stellar occultations to find rings around Neptune. These met with limited success in the mid-1980s. About 10 percent of the observed stellar occultations showed starlight blockage on one side of Neptune or the other, but never on both sides. This gave rise to the speculation that Neptune had a series of partial rings (ring arcs) occupying 10 percent or less of their orbit(s). Voyager 2 showed these ring arcs to be denser collections of ring particles in Neptune's continuous outermost ring, as discussed above. (E.D.M.)

Pluto

Pluto, designated ♇ in astronomy, is normally the outermost planet of the solar system and, by far, the smallest in size and mass. It is also the most recently discovered planet, having been found in 1930. It is not visible in the night sky to the unaided eye. Pluto's single moon, Charon, is close enough in size to Pluto that it has become common to refer to the two bodies as a double planetary system.

Pluto is named for the god of the underworld in Roman mythology (the Greek equivalent is Hades). Because of the planet's remoteness and small size, the best telescopes on Earth and in Earth orbit have been able to resolve little detail on its surface. Most of what is known about it has been learned since the late 1970s as an outcome of the discovery of Charon. In the 1990s, Pluto's very status as a planet was debated when new discoveries showed it to be a giant leftover of the process that built the solar system's other outer planets.

BASIC ASTRONOMICAL DATA

Pluto's mean distance from the Sun, about 5.9 billion kilometres (39.5 astronomical units), gives it the largest orbit among the planets. (One astronomical unit [AU] is the average Earth–Sun distance—about 150 million kilometres.)

Orbital eccentricity

Its orbit is atypical in several ways. It is the most elongated, or eccentric, of all the planetary orbits and the most inclined (at 17.1°) to the ecliptic, the plane of Earth's orbit, near which the orbits of most of the other planets lie. In traveling its eccentric path around the Sun, Pluto varies in distance from 29.7 AU (perihelion) to 49.5 AU (aphelion). Because Neptune, the next planet inward from Pluto, orbits in a nearly circular path at 30.1 AU, Pluto is for a small part of each revolution actually closer to the Sun than is Neptune. Nevertheless, the two planets will never collide, because Pluto is locked in a stabilizing 3:2 resonance with Neptune—*i.e.*, it completes two orbits around the Sun in exactly the time it takes Neptune to complete three. This gravitational interaction affects their orbits such that they can never pass closer than about 17 AU. The last time Pluto reached perihelion occurred in 1989; for about 10 years prior and again afterward, Neptune held the title of the most distant planet.

Eliot Young, Southwest Research Institute, NASA's Planetary Astronomy Program

Figure 57: Two-colour image of Pluto, created from telescopic data collected between 1985 and 1990 during a period of mutual eclipses of Pluto and its moon, Charon. Pluto's slightly reddish hue indicates that its surface does not comprise pure ices, though the nature of the coloured material remains to be determined.

Observations from Earth have revealed that Pluto's brightness varies with a period of 6.3873 Earth days, now well established as its rotation period. Only Mercury, with almost 59 days, and Venus, with 243 days, turn more slowly. Pluto's spin axis is tilted by 120° from the perpendicular to the plane of its orbit, so that the planet's north pole actually points 30° below the plane. (By convention, *above the plane* is taken to mean in the direction of Earth's and the Sun's north poles; *below*, in the opposite direction.) Pluto thus rotates nearly on its side in a retrograde direction (opposite the direction of rotation of the Sun and most of the planets); an observer on its surface would see the Sun rise in the west and set in the east.

Tilt of spin axis

The planet's anomalies also extend to its physical characteristics. Pluto's radius is only about two-thirds that of Earth's Moon. Next to the other outer planets—the giants Jupiter, Saturn, Uranus, and Neptune—it is strikingly tiny. When these characteristics are combined with what is known about its density and composition, Pluto appears to have more in common with the large icy moons of the outer planets than with any of the planets themselves. Its closest twin is Neptune's moon Triton, which suggests a similar origin for these two bodies (see below *Origin of Pluto and Charon*). For additional orbital and physical data about Pluto and Charon, see Table 23.

Table 23: Data for Pluto and Charon

Mean distance from Sun	5.91×10^9 km (39.5 AU)
Plutonian year (sidereal period of revolution)	247.69 Earth years
Eccentricity of orbit	0.251
Inclination of orbit to ecliptic	17.1°
Radius	1,172 km
Mass	1.2×10^{22} kg
Mean density	about 2 g/cm³
Rotation period (Plutonian sidereal day)	6.3873 Earth days
Mean synodic period	366.74 Earth days
Inclination of equator to orbit (obliquity)	120°
Moon	Charon
Mean distance from centre of Pluto (orbital radius)	19,640 km
Orbital period (sidereal period)	6.3873 Earth days
Rotation period	same as orbital period (synchronous)
Radius	625 km
Mass	1.8×10^{21} kg
Mean density	about 1.7 g/cm³

THE ATMOSPHERE

Although the detection of methane ice on Pluto's surface in the 1970s (see below *The surface and interior*) gave scientists confidence that the planet had an atmosphere, direct observation of it had to wait until the next decade. This confirmation was made in 1988 when Pluto passed in front of (occulted) a star as observed from Earth. The star's light gradually dimmed just before it disappeared behind the planet, demonstrating the presence of a thin, greatly distended atmosphere. Because Pluto's atmosphere must consist of vapours in equilibrium with their ices, small changes in temperature should have a large effect on the amount of gas in the atmosphere. During the years surrounding Pluto's perihelion (as in the period centred on 1989), when the planet is slightly less cold than average, more of its frozen gases will vapourize; the atmosphere should then be near or at its thickest. Astronomers in the year 2000 estimated a surface pressure in the range of a few to several tens of microbars (millionths of sea-level pressure on Earth). At aphelion, when Pluto is receiving the least sunlight, its atmosphere may not be detectable at all.

Variation of atmospheric pressure with solar distance

Drawing from occultation data and making reasonable assumptions about the atmospheric temperature, scientists have calculated that each particle—*i.e.*, each atom or mol-

ecule—of Pluto's atmosphere has a mean molecular weight of about 25 atomic mass units. This implies that significant amounts of gases heavier than methane (molecular weight 16) must also be present. Molecular nitrogen (molecular weight 28) must in fact be the dominant constituent, because nitrogen ice was discovered on the surface (see below *The surface and interior*) and is more volatile than methane ice. Nitrogen is also the main component of the atmospheres of both Triton and Saturn's largest satellite, Titan, as well as of Earth.

Although ongoing Earth-based observations will add to knowledge about the atmosphere and other aspects of the planet, major new insights will likely require a spacecraft visit. In the first years of the 21st century, scientists looked to the U.S. New Horizons spacecraft mission to Pluto, Charon, and the solar system beyond to provide much of the needed data. The mission plan called for a nine-year flight to the Pluto-Charon system followed by a 150-day flyby for investigation of the surfaces, atmospheres, interiors, and space environment of the two bodies.

THE SURFACE AND INTERIOR

Observations of Pluto show that its colour is slightly reddish, although much less red than Mars or Jupiter's moon Io. Thus the surface of Pluto cannot be composed simply of pure ices, a conclusion supported by the observed variation in brightness caused by the planet's rotation. Its average reflectivity, or albedo, is 0.55 (*i.e.*, it returns 55 percent of the light that strikes it), compared with 0.1 for the Moon and 0.8 for Triton.

Surface constituents

The first crude infrared spectroscopic measurements, made in 1976, revealed the presence of solid methane on Pluto's surface. Using new Earth-based instrumentation available in the early 1990s, observers discovered ices of water, carbon monoxide, and molecular nitrogen. Although nitrogen's spectral signature is intrinsically very weak, it is now clear that this substance must be the dominant surface constituent. The nature of the dark, reddish material remains to be determined; some mixture of organic compounds produced by photochemical reactions in atmospheric gases or surface ices seems a likely possibility. Brightness fluctuations observed during periods when Pluto and Charon mutually eclipse one another (see below *Pluto's moon*) reveal that the planet's south polar region is unusually bright. Brightness maps based on observations with the Earth-orbiting Hubble Space Telescope also reveal some of this heterogeneity, but only visiting spacecraft can provide the spatial resolution needed to make associations between brightness and surface composition or topography.

The same Pluto-Charon eclipses have allowed astronomers to estimate the masses and radii of the two bodies. From this information their densities have been calculated to fall between 1.92 and 2.06 grams per cubic centimetre for Pluto and between 1.51 and 1.81 grams per cubic centimetre for Charon. These values suggest that a significant fraction of each body consists of materials such as silicate rock and organic compounds denser than water ice (density of 1 gram per cubic centimetre). In analogy with the icy moons of Jupiter and Saturn, it is customary to assume that Pluto has a rocky core surrounded by a thick mantle of water ice. The frozen nitrogen, carbon monoxide, and methane observed on its surface are expected to be in the form of a relatively thin layer.

The surface temperature of Pluto has proved very difficult to measure. Observations made in 1983 from the Earth-orbiting Infrared Astronomical Satellite (IRAS) suggest values in the range of 45 to 58 K (−379° to −355° F, −228° to −215° C), whereas measurements from Earth's surface imply a slightly lower range of 35 to 50 K (−397° to −370° F, −238° to −223° C). The temperature certainly must vary over the surface, depending on the reflectivity at a given location and the angle of the noon Sun there. The solar energy falling on Pluto is also expected to decrease by a factor of roughly three as the planet moves from perihelion to aphelion.

Photo AURA/STScI/NASA/JPL

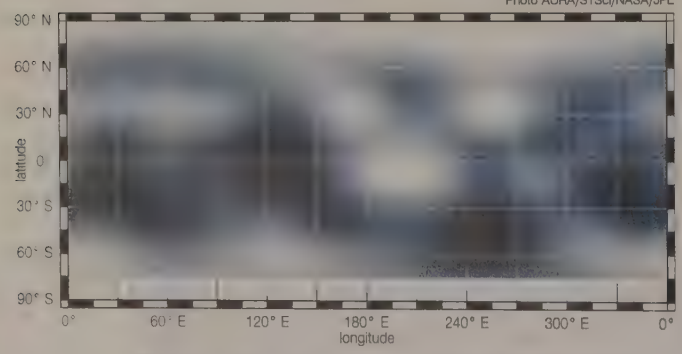


Figure 58: Map of Pluto's surface, a Mercator projection based on images made by the Hubble Space Telescope in June and July 1994. The north polar region generally comprises bright areas, while the equatorial region, particularly to the south, has more dark patches. The variations in brightness may indicate topographic features such as basins or craters; ground cover such as frost, rock, or dust; or a combination. Blue tint has been added in reproduction.

PLUTO'S MOON

Charon, which is fully half the size of Pluto, is the largest moon with respect to its primary planet in the solar system. It revolves around Pluto—more accurately, the two bodies revolve around a common centre of mass—at a distance of about 19,640 kilometres. Charon's period of revolution is exactly equal to the rotation period of the planet itself. In other words, Charon is in synchronous orbit around Pluto, the only moon in the solar system to have that distinction. As a result, Charon remains above the same location on Pluto's surface, never rising or setting (just as do communication satellites in geostationary orbits over Earth). In addition, as with most moons in the solar system, Charon is in a state of synchronous rotation—*i.e.*, it always presents the same face to its primary planet.

Charon is somewhat less reflective (has a lower albedo—about 0.35) than Pluto and is more neutral in colour. Its spectrum reveals the presence of water ice, which appears to be the dominant surface constituent. There is no hint of the solid methane that is so obvious on its larger neighbour. As discussed above, Charon's density implies that the moon contains materials such as silicates and organic compounds that are denser than water ice. Through an extraordinary coincidence, in 1985—just seven years after Charon was discovered—it began a five-year period of mutual eclipse events with Pluto in which the moon alternately crossed the disk of and was hidden by Pluto, as seen from Earth, every 6.4 days (see Figure 59). These events occur when Earth passes through Charon's orbital plane around Pluto, which happens only twice in Pluto's 248-year orbit. Careful observations of these events allowed more precise determinations of the radii and masses of Pluto and Charon than heretofore possible. They also permitted astronomers to estimate individual overall albedos of the two bodies and even to create surface maps depicting brightness differences.

Pluto-Charon mutual eclipse events

DISCOVERIES OF PLUTO AND CHARON

Pluto was the third planet to be discovered, after Uranus and Neptune, as opposed to the six planets that have been visible to the naked eye since ancient times. Its existence had been postulated since the late 19th century on the basis of apparent perturbations of the orbital motion of Uranus, which suggested that a more distant planet was gravitationally disturbing it. Astronomers later realized that these perturbations were spurious—Pluto's mass is too small to have caused the suspected disturbances. Thus, Pluto's discovery was a remarkable coincidence attributable to careful observations rather than to accurate prediction of the existence of a hypothetical planet.

The search for the expected ninth planet was supported most actively at the Lowell Observatory in Flagstaff, Arizona, U.S., in the early 20th century. It was initiated by the observatory's founder, Percival Lowell, an American astronomer who had achieved notoriety through his highly publicized claims of canal sightings on Mars. After two attempts to find the planet prior to Lowell's death in 1916, an astronomical camera built for the purpose and capable

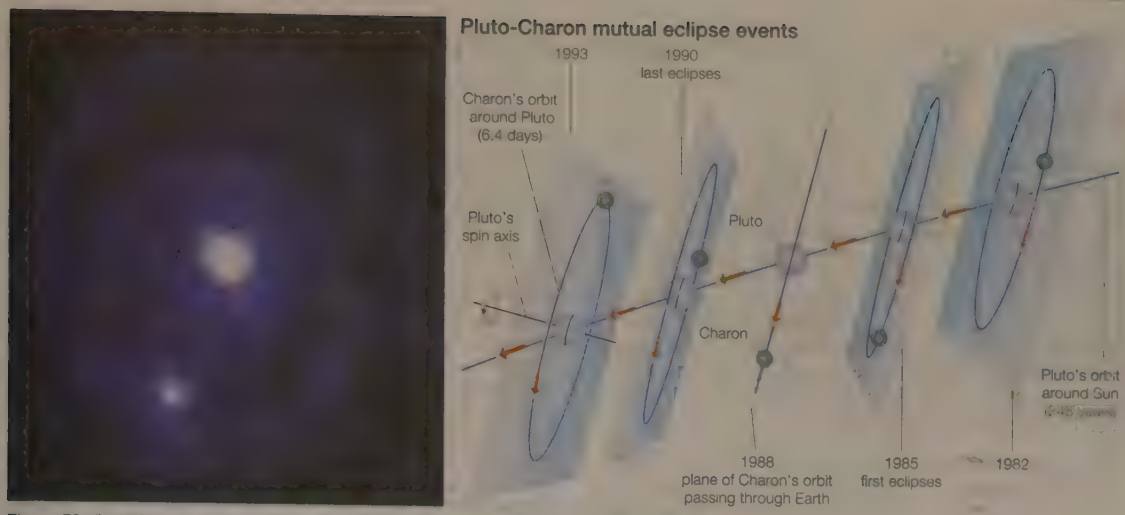


Figure 59: (Left) Pluto and Charon (centre and lower left, respectively) in an image made in 1990 by the Hubble Space Telescope. (Right) Charon's orbit around Pluto as viewed from Earth, 1982–93. Between 1985 and 1990, Pluto and Charon were in a period of mutual eclipses. During each 6.4-day revolution around Pluto, Charon passed in front of the planet, partially blocking it from view, and then disappeared behind the planet.
(Left) From NASA/European Space Agency; (right) Encyclopædia Britannica, Inc

of collecting light from a wide field of sky was put into service in 1929, and a young amateur astronomer, Clyde Tombaugh, was hired to carry out the search. On Feb. 18, 1930, less than a year after he began his work, Tombaugh found Pluto in the constellation Gemini. The object appeared as a dim 15th-magnitude "star" that slowly changed position against the fixed background stars as it pursued its 248-year orbit around the Sun. The symbol invented for it, ♇, stands both for the first two letters of Pluto and for the initials of Percival Lowell.

Charon's discovery

Charon was discovered in 1978 in images of Pluto that had been recorded photographically at the U.S. Naval Observatory station in Flagstaff, only a short distance from the site of Pluto's discovery. These images were being made by James W. Christy and Robert S. Harrington in an attempt to obtain more accurate measurements of Pluto's orbit. The new moon was named after the boatman in Greek mythology who ferries dead souls to Hades' realm in the underworld.

Prior to the discovery of Charon, Pluto was thought to be larger and more massive than it actually is. Even in the discovery images, Charon appears as an unresolved bump on the side of Pluto, an indication of the observational difficulties involved. Only near the end of the 20th century, with the availability of the Hubble Space Telescope and Earth-based instruments equipped with adaptive optics that compensate for atmospheric turbulence, did astronomers first resolve Pluto and Charon into separate bodies.

ORIGIN OF PLUTO AND CHARON

Before the discovery of Charon, it was popular to assume that Pluto was a former moon of Neptune that had somehow escaped its orbit. This idea gained support from the apparent similarity of the dimensions of Pluto and Triton and the near coincidence in Triton's orbital period (5.9 days) and Pluto's rotation period (6.4 days). It was suggested that a close encounter between these two bodies when they were both moons led to the ejection of Pluto from the Neptunian system and caused Triton to assume the retrograde orbit that is presently observed.

Astronomers found it difficult to establish the likelihood that all these events would have occurred, and the discovery of Charon provided information that further refuted the theory. Because the revised mass of Pluto is only half that of Triton, Pluto clearly could not have caused the reversal of Triton's orbit. Also, the fact that Pluto has a proportionally large moon of its own makes the escape idea implausible. Current thinking favours the idea that Pluto and Charon formed as two independent bodies in the solar nebula, the gaseous cloud from which the solar system condensed (see above *An overview of the solar system: Origin of the solar system*). A collision between Pluto and a proto-Charon could have produced a debris ring around Pluto that accreted by gravitational attraction to form the present moon. This scenario is similar to the currently favoured model for the formation of Earth's Moon (see above *The Moon: Origin and evolution*).

This scenario implies that at the time the Pluto-Charon system formed, about 4.6 billion years ago, the outer solar nebula contained many icy bodies of about the same size as these two. The bodies themselves are thought to have been built up from smaller entities that today would be recognized as the nuclei of comets. Triton is presumably another of these large icy planetesimals, captured into orbit by Neptune in the planet's early history.

Pluto as a Kuiper belt object

Most of these icy planetesimals were incorporated into the cores of the giant planets during their formation. Many others, however, are thought to have remained as unconsolidated debris to make up the Kuiper belt—a disk-shaped region that lies beyond Neptune's orbit and, significantly, includes the outer part of Pluto's orbit. After hundreds of Kuiper belt objects (KBOs) were directly observed starting in the early 1990s, astronomers came to suspect that Pluto (with Charon) is the largest known member of the Kuiper belt and that bodies such as Triton and some other icy moons of the outer planets originated as KBOs. In fact, like Pluto, one group of KBOs have highly eccentric orbits inclined to the plane of the solar system, and they exhibit the same 3:2 orbital resonance with Neptune. In recognition of this affinity, these objects have been dubbed Plutinos ("little Plutos").

PLUTO'S STATUS AS A PLANET

If Pluto had been discovered in the context of the Kuiper belt, rather than as an isolated entity, it might not have been ranked with the other eight planets. Furthermore, it may be only a matter of time before a body even larger than Pluto is detected lurking farther out in the Kuiper belt. In fact, within just a few years surrounding the turn of the 21st century, several KBOs each the size of Charon were discovered, including one that is well outside the gravitational influence of Neptune.

Nonetheless, since the early 1990s the major scientific discussions over Pluto's status have concluded with most of the participants agreeing that Pluto should remain a planet. One of these took place in 2000 at the General Assembly of the International Astronomical Union, the organization charged with classifying astronomical objects and that originally classified Pluto. Pluto's planetary status has been accepted by scientists and the general public for the greater part of a century. A demotion does not appear likely to be in its future. (T.C.O.)

OTHER CONSTITUENTS OF THE SOLAR SYSTEM

Asteroids

Asteroids are rocky bodies about 1,000 kilometres or less in diameter that orbit the Sun primarily between the orbits of Mars and Jupiter. Because of their small size and large numbers relative to the nine major planets, asteroids are also called minor planets. The two designations are frequently used interchangeably, though dynamicists, astronomers who study individual objects with dynamically interesting orbits or groups of objects with similar orbital characteristics, generally use the term minor planet, whereas those who study the physical properties of such objects usually refer to them as asteroids. The term planetoid is sometimes used as well.

Nomenclature

HISTORICAL SURVEY OF MAJOR ASTEROID DISCOVERIES

Early observations. The first asteroid was discovered on Jan. 1, 1801, by Giuseppe Piazzi at Palermo, Italy. At first Piazzi thought that he had discovered a comet; however, after the orbital elements of the object had been computed it became clear that the object moved in a planetlike orbit between the orbits of Mars and Jupiter. Owing to illness, Piazzi was able to observe the object only until February 11, and, as no one else was aware of its existence, it was not reobserved before it moved into the daytime sky. The short arc of observations did not allow computation of an orbit of sufficient accuracy to predict where the object would reappear when it moved back into the night sky, and so it was "lost." There matters might have stood were it not for the fact that this object was located at the heliocentric distance predicted by Bode's law of planetary distances proposed in 1766 by the German astronomer Johann D. Titius and popularized by his compatriot Johann E. Bode, who used the scheme to advance the notion of a "missing" planet between Mars and Jupiter. The discovery of Uranus in 1781 by the British astronomer William Herschel at a distance that closely fit the distance predicted by Bode's law was taken as strong evidence of its correctness. Some astronomers were so convinced that during an astronomical conference in 1796 they agreed to undertake a systematic search. Ironically, Piazzi was not a party to this attempt to locate the missing planet. Nonetheless, Bode and others, on the basis of the preliminary orbit, believed that Piazzi had found and then lost it. This led the German mathematician Carl Friedrich Gauss to develop in 1801 a method for computing the orbit of an asteroid from only a few observations, a technique that has not been significantly improved since. Using Gauss's predictions, the German astronomer Franz von Zach rediscovered Ceres on Jan. 1, 1802. Piazzi named this object Ceres after the ancient Roman grain goddess and patron goddess of Sicily, thereby initiating a tradition that continues to the present day: asteroids are named by their discoverers (in contrast to comets, which are named for their discoverers).

Identification of Ceres

The discovery of three more faint (compared with Mars and Jupiter) objects in similar orbits over the next six years (Pallas, Juno, and Vesta, respectively) complicated this elegant solution to the missing-planet problem and gave rise to the surprisingly long-lived, though no longer generally accepted, idea that the asteroids were remnants of a planet that had exploded.

Following this flurry of activity, the search for the planet appears to have been abandoned until 1830, when Karl L. Hencke renewed it. In 1845 he discovered the fifth asteroid, which he named Astraea.

Modern research. There were 88 known asteroids by 1866, when the next major discovery was made: Daniel Kirkwood, an American astronomer, noted that there were gaps in the distribution of asteroid distances from the Sun (see below *Distribution and Kirkwood gaps*). The introduction of photography to the search for new asteroids by the German astronomer Max Wolf in 1891, by which time 322 asteroids had been identified, accelerated the discovery rate. By the end of the 19th century, 464 had been found. The asteroid designated 323 Brucia, detected by Wolf in 1891, was the first to be discovered by means of photography.

Application of photography

The first measurements of the sizes of asteroids were made in 1894 and 1895 by the American astronomer Edward E. Barnard, who used a filar micrometer (an instrument normally employed for visual measurement of the separations of double stars) to estimate the diameters of the first four asteroids. Barnard's results established that Ceres was the largest asteroid, with an estimated diameter of nearly 800 kilometres. These values remained the best available until new techniques were introduced during the early 1970s (see below *Size and albedo*). The first four asteroids came to be known as "the big four," and, because all other asteroids were much fainter, they were believed to be considerably smaller as well.

In 1918 the Japanese astronomer Kiyotsugu Hirayama recognized clustering in three of the orbital elements of various asteroids (semimajor axis, eccentricity, and inclination). He speculated that objects sharing these elements had been formed by explosions of larger parent asteroids and called such groups of asteroids "families."

The idea that Jupiter was responsible for interrupting the formation of a planet from the swarm of planetesimals accreting near a heliocentric distance of 2.8 AU was introduced in 1944 by O.J. Schmidt. In 1951 the Estonian astronomer Ernest J. Öpik calculated the lifetimes of asteroids with orbits that passed close to those of the major planets and showed that most such asteroids were destined to collide with a planet or be ejected from the solar system on time scales of a few hundred thousand to a few million years. Since the age of the solar system is approximately 4.6 billion years, this meant that the asteroids seen today in such orbits must have entered them recently and implied that there was a source for the asteroids. Öpik believed this source to be comets that had been captured by the planets and that had lost their volatile material through repeated passages inside the orbit of Mars.

The mass of Vesta was deduced by the German-born American astronomer Hans G. Hertz in 1966 from measurements of its perturbations on the orbit of 197 Arete. The first mineralogical determination of the surface composition of an asteroid was made in 1969 by Thomas McCord, John B. Adams, and Torrence V. Johnson of the United States, who used spectrophotometry to identify the mineral pyroxene in the surface material of 4 Vesta. In 1970 the first reliable albedos (reflectivities) and diameters of asteroids were determined by two groups of American astronomers—Joseph F. Veverka, Benjamin H. Zellner, and their colleagues, who used a technique based on polarization measurements, and David A. Allen and Dennis L. Matson, who employed infrared radiometry.

Development of classification systems. In 1975 Zellner, together with Clark R. Chapman and David D. Morrison, grouped the asteroids into three broad taxonomic classes, which they designated C, S, and M (see below *Composition*). They estimated that about 75 percent belonged to class C, 15 percent to class S, and 5 percent to class M. The remaining 5 percent were unclassifiable in their system owing to either poor data or genuinely unusual properties. Furthermore, they noted that the S class dominated the population at the inner edge of the asteroid belt, whereas the C class was dominant in the middle and outer regions of the belt. In 1982 other American astronomers, Jonathan C. Gradie and Edward F. Tedesco, expanded this taxonomic system and recognized that the asteroid belt consisted of rings of differing taxonomic classes with the S, C, P, and D classes dominating the populations at heliocentric distances of approximately 2, 3, 4, and 5 AU, respectively (see below *Composition;* see also Figure 57). As more data became available from further observations, additional minor classes were recognized by the American astronomer David J. Tholen in 1984, by the Italian astronomer Antonietta Barucci and colleagues in 1987, and by Tedesco and colleagues in 1989.

ORBITS OF ASTEROIDS

Asteroid designations

Because of their large numbers, asteroids are assigned numbers as well as names. The numbers are assigned consecutively after accurate orbital elements have been determined. (For example, Ceres is officially known as 1 Ceres, Pallas as 2 Pallas, and so forth.) By mid-1991, more than 7,000 asteroids had been observed at two or more oppositions, and 5,000 of these were numbered. The discoverers have the right to choose a name for their discoveries as soon as they are numbered. Now the names selected are submitted to the International Astronomical Union for approval.

Prior to the mid-20th century, asteroids were sometimes assigned numbers before accurate orbital elements had been determined, and so some numbered asteroids could not later be located. These objects are referred to as "lost" asteroids. As of 1991 only asteroid 719 Albert remained lost.

The Minor Planet Center at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Mass., U.S., maintains computer files for all measurements of asteroid positions. The Institute for Theoretical Astronomy in St. Petersburg, Russia, publishes each year the *Ephemerides of Minor Planets,* which contains the orbital elements of all numbered asteroids, together with their opposition dates and ephemerides.

Although most asteroids travel in fairly circular orbits, there are some notable exceptions. One of the most extreme of these is 3200 Phaethon, discovered by the Infrared Astronomical Satellite (IRAS) in 1983. (It was the first asteroid to be discovered by a spacecraft.) Phaethon approaches to within 0.14 AU of the Sun, well within Mercury's perihelion distance of 0.31 AU. Phaethon's aphelion (2.4 AU) is in the main asteroid belt. This object is the parent body of the Geminid meteor stream and, since the parent bodies of all other meteor streams identified to date are comets, it is considered by some to be a defunct comet. Another asteroid, 944 Hidalgo, is also thought by some to be a defunct comet because of its unusual orbit. This object, discovered in 1920 by Walter Baade at the Bergedorf Observatory near Hamburg, Ger., has a perihelion distance of 2.02 AU at the inner edge of the main asteroid belt and an aphelion distance of 9.68 AU just beyond the orbit of Saturn at 9.54 AU. Finally, there is the case of 2060 Chiron, discovered in 1977 by Charles Kowal at the Palomar Observatory near San Diego, Calif., U.S. This object was originally classified as an asteroid, but in 1989 the American astronomers Karen J. Meech and Michael J. Belton observed a dusty coma surrounding it, and in 1991 Schelte J. Bus, also of the United States, and his colleagues detected the presence of cyanogen radicals, a known constituent of the gas comas of comets. Chiron travels in an orbit that lies wholly outside of the asteroid belt, having a perihelion distance of 8.43 AU (between the orbits of Jupiter and Saturn) and an aphelion distance of 18.8 AU, which nearly reaches the orbit of Uranus at 19.2 AU. Because Chiron moves in a chaotic, planet-orbit-crossing orbit, astronomers believe that it will eventually collide with a planet or be permanently ejected from the solar system (see also below *Groups of comets and other unusual cometary objects*).

Distribution and Kirkwood gaps. About 95 percent of the known asteroids move in orbits between those of Mars and Jupiter. These orbits, however, are not uniformly distributed but, as shown in Figure 60, exhibit "gaps" in the distribution of their semimajor axes. These so-called Kirkwood gaps are due to resonances with Jupiter's orbital period. An asteroid with a semimajor axis of 3.3 AU, for example, makes two circuits around the Sun in the time it takes Jupiter to make one and is thus said to be in a two-to-one (written 2:1) resonance orbit with Jupiter. Consequently, once every two orbits Jupiter and an asteroid in such an orbit would be in the same relative positions, and such an asteroid would experience a force in a fixed direction. Repeated applications of this force would eventually change the semimajor axes of asteroids in such orbits, thus creating a gap at that distance. Gaps occur at 4:1, 7:2, 3:1, 5:2, 7:3, and 2:1 resonances, while concentrations occur at the 3:2 (Hilda group), 4:3 (Thule), and 1:1 (Trojan group) resonances. The presence of secular resonances complicates the situation, particularly at the inner edges of the belt. An adequate explanation of why some resonances produce gaps and others produce concentrations has yet to be found.

Cause of the Kirkwood gaps

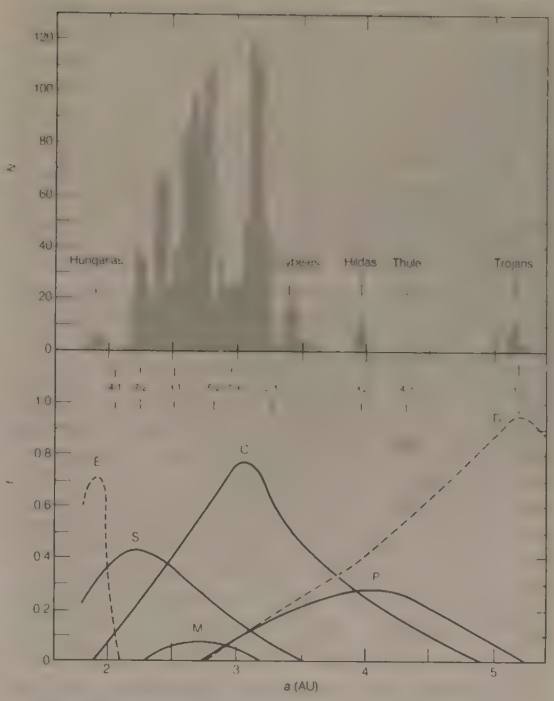


Figure 60: (Top) Number distribution (*N*) in semimajor axis (*a*) for asteroids with diameters greater than about 50 kilometres. The fractions below the centre indicate the locations of the major orbital period resonances with Jupiter. The deep Kirkwood gaps occur within the main belt near the 3:1, 5:2, and 7:3 resonances, whereas the concentrations are found at the 3:2, 4:3, and 1:1 resonances. The main belt is confined to the region between the 4:1 and 2:1 resonances. The Kirkwood gap at the 7:2 resonance is too narrow to be displayed at the scale plotted. (Bottom) Fractional number distribution (*f*) of the major asteroid classes from Table 24. The sum of the various classes found at any given semimajor axis must equal 1. The other classes in Table 24 do not appear because none of them constitute more than 2 percent of the total number of asteroids found at any given semimajor axis.

From J. Gradie and E. Tedesco, *Science*, vol. 216, p. 1406 (June 25, 1982). © American Association for the Advancement of Science

Families. Within the main asteroid belt are groups of asteroids that cluster in certain orbital elements (semimajor axis, eccentricity, and inclination). Such groups are called families and assigned the name of the lowest numbered asteroid in the family. Asteroid families are thought to be formed when an asteroid is disrupted in a catastrophic collision. Theoretical studies indicate that such catastrophic collisions between asteroids are common enough to account for the number of families observed. About 40 percent of all known asteroids belong to such families.

The three largest families (Eos, Koronis, and Themis) have been determined to be compositionally homogeneous. If the asteroids belonging to them are considered to be fragments of a single parent body, then their parent bodies must have had diameters of 200, 90, and 300 kilometres, respectively. The smaller families have not been as well studied because their numbered members are fewer and smaller (and hence fainter). Nevertheless, it is known that some of the smaller families are compositionally inhomogeneous and that, at least in some cases, what are observed are pieces of a geochemically differentiated parent body. It is theorized that some of the Earth-crossing asteroids and meteorites reaching the terrestrial surface are fragments produced in collisions similar to those that produced the asteroid families.

Outer-belt asteroids. The vast majority of asteroids have orbital periods between three years and six years—*i.e.*, between one-fourth and one-half of Jupiter's orbital period (see Figure 60). These asteroids are said to be

Asteroid groups that fall outside the main belt

main-belt asteroids. Besides the few asteroids in highly unusual orbits, some of which were noted above, there are a number of groups that fall outside the main belt. Those that have orbital periods greater than one-half that of Jupiter are called outer-belt asteroids. There are four such groups: the Cybeles, Hildas, and Thule, named after the lowest numbered asteroid in each group, as well as the Trojan group, so called because all its members are named after characters from Homer's epic work about the Trojan War, the *Iliad*. As of mid-1991, 86 Cybeles, 66 Hildas, 1 Thule, and 153 Trojans had been observed at two or more oppositions.

Trojan asteroids. In 1772 the French mathematician and astronomer Joseph-Louis Lagrange predicted the existence and location of two groups of asteroids located near the (L_4 and L_5) equilateral triangular stability points of a three-body system formed by the Sun, Jupiter, and the asteroids. These are two of the five stable points in the circular, restricted three-body problem. (The other three stable points are located along a line passing through the Sun and Jupiter. Because of the presence of other planets, principally Saturn, the Sun-Jupiter-Trojan asteroid system is not a true three-body system, and so these other three points are not stable and no asteroids have been found near them. In fact, most of Jupiter's Trojan asteroids do not move in the plane of its orbit but rather in orbits inclined by up to 40° and at longitudes that differ by as much as 70° from the longitudes of the true Lagrangian points.)

In 1906 Max Wolf discovered 588 Achilles near the Lagrangian point preceding Jupiter in its orbit. Within a year August Kopff had discovered two more: 617 Patroclus, located near the following Lagrangian point, and 624 Hector near the preceding Lagrangian point. It was later decided to name such asteroids after the participants in the Trojan War as given in the *Iliad* and, furthermore, to name those near the preceding point after Greek warriors and those near the following point after Trojan warriors. With the exception of the two previously named "spies" (Hector, the lone Trojan in the Greek camp, and Patroclus, the lone Greek in the Trojan camp), this tradition has been maintained.

The term Trojan has since been applied to any object occupying the equilateral Lagrangian points of other pairs of bodies. Searches have been made for Trojans of the Earth, Saturn, and Neptune, as well as of the Earth-Moon system, but so far none have been found. It was long considered doubtful whether truly stable orbits could exist near these Lagrangian points because of perturbations by the major planets. However, in 1990 an asteroid librating about the following Lagrangian point of Mars was discovered by the American astronomers David H. Levy and Henry E. Holt, thus reopening this question.

Although only about 65 Trojans have been numbered, photographic surveys have shown that there are about $2{,}300 \pm 500$ such asteroids with diameters greater than 15 kilometres. About 1,300 of these are located near the preceding Lagrangian point and 1,000 near the following Lagrangian point.

The Hungaria group

Inner-belt asteroids. There is only one known group of inner-belt asteroids—namely, the Hungarias. The Hungaria asteroids, about 100 of which are known, have a mean semimajor axis of 1.91 AU; thus, their orbital periods are less than one-fourth that of Jupiter (see Figure 57). Hungarias have nearly circular orbits (the mean eccentricity is 0.08) but large inclinations, the mean inclination being 22.7°. At least one dynamical family, the Hungaria family, exists within the Hungaria group. Because of the low eccentricities of their orbits, the mean perihelion distance of a Hungaria asteroid is 1.76 AU. Accordingly, a typical Hungaria cannot pass close to Mars, whose aphelion distance is 1.67 AU. A few Hungarias, however, have perihelion distances a few hundredths of 1 AU less than Mars's aphelion distance and so are shallow "Mars crossers" (see below) as well.

Classes of near-Earth asteroids

Near-Earth asteroids. Asteroids that can pass inside the orbit of Mars are said to be near-Earth asteroids. The near-Earth asteroids are subdivided into several classes. The most distant—those that can cross the orbit of Mars but that have perihelion distances (q) greater than 1.3 AU—are dubbed Mars crossers. This group is further subdivided into two groups: shallow Mars crossers ($1.58 \leq q < 1.67$ AU) and deep Mars crossers ($1.3 < q < 1.58$ AU).

The next most distant group of near-Earth asteroids are the Amors ($1.017 < q \leq 1.3$ AU). Amor asteroids have perihelion distances greater than the Earth's present aphelion distance (Q) of 1.017 AU and therefore do not at present cross the planet's orbit. Because of strong gravitational perturbations produced by their close approaches to the Earth, however, the orbital elements of all the Earth-approaching asteroids, except the shallow Mars crossers, change appreciably on a time scale of a few years or tens of years. For this reason about half the known Amors, including 1221 Amor, are part-time Earth crossers. Only asteroids that cross the orbits of planets, such as the Earth-approaching asteroids and objects like 944 Hidalgo and 2060 Chiron, suffer significant changes in their orbital elements on time scales shorter than many millions of years. Hence, the outer-belt asteroid groups (Cybeles, Hildas, Thule, and the Trojans) do not interchange members.

There are two groups of near-Earth asteroids that deeply cross the Earth's orbit on an almost continuous basis. The first of these to be discovered were the Apollo asteroids, 1862 Apollo being detected by the German astronomer Karl Wilhelm Reinmuth in 1932 but lost shortly thereafter and not rediscovered until 1978. Apollo asteroids have semimajor axes (a) that are greater than or equal to 1 AU and perihelion distances that are less than or equal to 1.017 AU; thus, they cross the Earth's orbit when near the perihelia of their orbits. For the other group of Earth-crossing asteroids—named Atens after 2062 Aten, which was discovered in 1976 by Eleanor F. Helin of the United States—$a < 1.0$ AU and $Q \geq 0.983$ AU, the present perihelion distance of the Earth. These asteroids cross the Earth's orbit when near the aphelia of their orbits.

By mid-1991 the number of known Aten, Apollo, and Amor asteroids were 11, 91, and 81, respectively. Most of these were discovered since 1970, when dedicated searches for this type of asteroid were begun: 25 were discovered during the 1970s, 80 during the 1980s, and 49 during the first 20 months of the 1990s. It is estimated that there are roughly 100 Atens, 700 Apollos, and 1,000 Amors that have diameters larger than about one kilometre. Because these asteroids travel in orbits that cross the Earth's orbit, close approaches to the Earth occur, as well as occasional collisions. For example, in January 1991, an Apollo asteroid with an estimated diameter of 10 metres passed by the Earth within less than half the distance to the Moon.

THE NATURE OF ASTEROIDS

Rotation and shape. Asteroid rotational periods and shapes are determined primarily by monitoring their changing brightness on time scales of hours to days. Short-period fluctuations in brightness caused by the rotation of an irregularly shaped or spotted body (a spotted body being a spherical object with albedo differences) give rise to a light curve (a graph of brightness versus time) that repeats at regular intervals corresponding to an asteroid's rotation period. The range of brightness variation is more difficult to interpret but is closely related to an asteroid's shape or spottedness.

Rotational periods have been determined for more than 400 asteroids. They range from 2.3 hours to 48 days, but the majority (more than 80 percent) lie between 4 hours and 20 hours. Periods longer than a few days may actually be due to precession caused by an unseen satellite. The mean rotational period is roughly 10 hours for the entire sample. The largest asteroids (those with diameters greater than about 175 kilometres), however, have a mean rotational period close to 7 hours, whereas this value is about 10 hours for smaller asteroids. The largest asteroids may have preserved their primordial rotation rates, but the smaller ones have almost certainly had theirs modified by subsequent collisions. The difference in rotation rates between the larger and the smaller asteroids is believed to stem from the fact that large asteroids retain all of their collisional debris from minor collisions, whereas smaller asteroids retain more of the debris ejected in the direction

opposite to that of their spins, causing a loss of angular momentum and thus a reduction in speed of rotation.

Major collisions can completely disrupt smaller asteroids. The debris from such collisions makes still smaller asteroids, which can have virtually any shape or spin rate. Thus, the fact that no rotational periods shorter than about 2.5 hours have been observed implies that the material of which asteroids are made is not strong enough to withstand the centripetal forces that such rapid spins would produce.

For mathematical reasons it is impossible to distinguish between the rotation of a spotted sphere and an irregular shape of uniform reflectivity on the basis of observed brightness changes alone. Nevertheless, the fact that opposite hemispheres of most asteroids appear to have albedos differing by no more than a few percent suggests that their brightness variations are due mainly to changes in their projected, illuminated, visible, cross-sectional areas. Hence, in the absence of evidence to the contrary, it is generally accepted that variations in reflectivity contribute little to the observed rotational light-curve amplitude. Asteroid 4 Vesta is a notable exception to this generalization because it is known that the difference in reflectivity between its opposite hemispheres is sufficient to account for much of its light-curve amplitude.

Asteroid light-curve amplitudes range from zero to a factor of 6.5 in the case of the Apollo asteroid 1620 Geographos. A light-curve amplitude of zero is caused by viewing an asteroid along one of its rotational poles, while the 6.5 to 1 variation in brightness is believed to result from either of two possibilities: Geographos is a cigar-shaped object that is viewed along a line perpendicular to its rotational axis, or it is a pair of objects nearly in contact that orbit each other around their centre of mass.

The mean rotational amplitude for asteroids is about a factor of 1.3. These data, together with the assumptions mentioned above, allow astronomers to estimate asteroid shapes, which occur in a wide range (see Figure 61). Some asteroids, such as 1 Ceres, 2 Pallas, and 4 Vesta, are nearly spherical, whereas others, like 15 Eunomia, 107 Camilla, and 511 Davida, are quite elongated. Still others, as, for example, 624 Hektor, 1580 Betulia (not shown in Figure 58 but whose proposed shape is that of a kidney bean), and 4769 Castalia (which appears in radar observations by the American astronomer Steven J. Ostro to resolve as two spheres in contact), apparently have bizarre shapes.

Size and albedo. The most widely used technique for determining the sizes of asteroids is that of thermal radiometry. This technique makes use of the fact that the infrared radiation (heat) emitted by an asteroid must balance the solar radiation it absorbs. By using a so-called thermal model to balance the intensity of infrared radiation with the intensity at visual wavelengths, investigators are able to arrive at the diameter of the asteroid. Several other techniques—polarimetry, speckle interferometry, and radar—also are used, but they are limited to brighter, larger, and/or closer asteroids.

Application of thermal radiometry

From K. Beatty, B. O'Leary, and A. Chaikin (eds.), *The New Solar System*, 2nd ed.

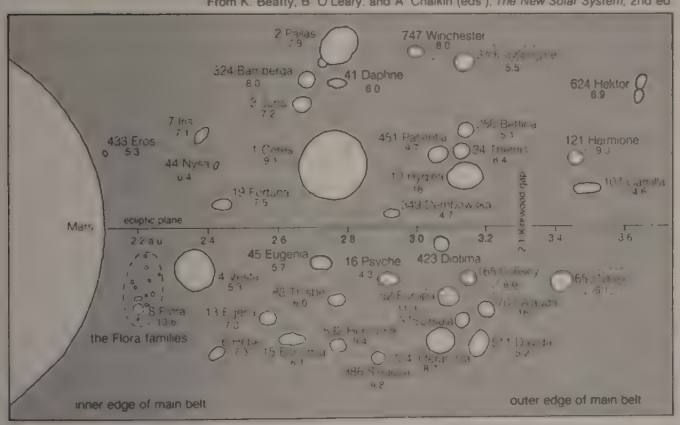


Figure 61: This representation of notable asteroids includes all those exceeding 200 kilometres in diameter. The asteroids are depicted in their correct relative sizes and shapes (the limb of Mars is included for comparison), and their rotational pole orientations (if known) are indicated. They are positioned at their correct relative distances from the Sun. Asteroids located near the top or bottom of the diagram travel in relatively eccentric and/or inclined orbits, whereas those near the ecliptic move in fairly circular, noninclined orbits. Rotational periods (in hours) are indicated beneath the names of the asteroids. Among the special smaller asteroids shown are the members of the Flora families larger than 15 kilometres across.

The only technique that measures the diameter directly (*i.e.*, without having to "model" the actual observations) is that of stellar occultation. In this method, investigators measure the length of time that a star disappears owing to the passage of an asteroid between the Earth and the star. Then, using the known distance and the rate of motion of the asteroid, they are able to determine the latter's diameter uniquely. The results of this method have made it possible to reliably calibrate the indirect techniques, thermal radiometry in particular. As a consequence, asteroid diameters obtained by means of such techniques are now thought to have uncertainties of less than 10 percent. Because passages of asteroids in front of stars are rare and best applied to fairly spherical asteroids and because only one cross section is measured, the majority of asteroid sizes have been obtained using indirect techniques.

By mid-1991 the diameters of about 21 asteroids, including 1 Ceres, 2 Pallas, 3 Juno, and 4 Vesta, had been accurately determined by means of the stellar-occultation method, whereas about 2,000 had been measured with the indirect techniques, principally thermal radiometry.

Asteroid 1 Ceres, with a diameter of about 930 kilometres (km), is the largest, followed by 2 Pallas at 535 km and 4 Vesta at 520 km. The fourth largest asteroid, 10 Hygiea, has a diameter of about 410 km. There are two asteroids with diameters between 300 and 400 km and about 24 with diameters between 200 and 300 km. In total, there are about 30 asteroids with diameters greater than 200 km. The smallest known asteroids are members of the Earth-approaching groups, since these asteroids can approach the Earth to within a few hundredths of 1 AU. The smallest routinely observed Earth-approaching asteroids measure less than 0.5 km across. It has been estimated that there are 250 asteroids larger than 100 km in diameter and perhaps 1,000,000 with diameters greater than 1 km.

Albedo

A parameter closely related to size is albedo, or reflectivity. This property also provides compositional information. Albedo is the ratio between the amount of light actually reflected and that which would be reflected by a uniformly scattering disk of the same size. Snow has an albedo of approximately 1 and coal an albedo of about 0.05.

An asteroid's apparent brightness depends on both its albedo and diameter as well as on its distance. For example, if 1 Ceres and 4 Vesta could both be observed at the same distance, Vesta would be the brighter of the two by about 15 percent. Vesta's diameter, however, is only about 55 percent that of Ceres. It appears brighter because its albedo is around 0.35, compared with 0.09 for Ceres.

Asteroid visual geometric albedos range from just under 0.02 to over 0.5 and may be divided into four albedo groups: low (0.02–0.07), intermediate (0.08–0.12), moderate (0.13–0.28), and high (greater than 0.28). About 78 percent of the known asteroids are low-albedo objects, and most of them are located in the outer half of the main asteroid belt and among the outer-belt populations. More than 95 percent of outer-belt asteroids belong to this group. Roughly 18 percent of all known asteroids belong to the moderate-albedo group, the vast majority of which are found in the inner half of the main belt. The high-albedo asteroids make up the remaining 4 percent of the asteroid population. For the most part, they occupy the same regions of the main asteroid belt as the moderate-albedo objects.

Mass and density. Asteroid masses are low and have little effect on the orbits of the major planets. In the past, the masses of several have been measured by noting their effect on the orbits of other asteroids that they approach closely at regular intervals. In 1989 the American scientists E. Myles Standish and Ronald W. Hellings used radio-ranging measurements that were transmitted from

the surface of Mars between June 1976 and August 1980 by the two Viking landers to determine distances to Mars with an accuracy of about 10 metres. Because the largest asteroids (Ceres, Pallas, and Vesta) cause perturbations in Mars's orbit in excess of 50 metres on time scales of 10 years or less, they were able to use the measured departures of Mars from its predicted orbit to estimate the masses (given below) of these three asteroids. Since the diameters of these three largest asteroids also have been determined and since their shapes are either spherical or ellipsoidal, their volumes are known as well. Knowledge of the mass and volume of an asteroid allows its density to be computed.

The mass of the largest asteroid, Ceres, is 1.0×10^{24} grams (0.0002 the mass of the Earth). The masses of the second and third largest asteroids, Pallas and Vesta, are only 0.28 and 0.30 times the mass of Ceres. The mass of the entire asteroid belt is roughly three times that of Ceres. Most of the mass in the asteroid belt is concentrated in the larger asteroids. The 10th largest asteroid has only 1/60 the mass of the largest, and about 90 percent of the total mass is contained in asteroids with diameters exceeding 100 kilometres. Ninety percent of the total mass of the asteroids is located in the main belt, 9 percent is in the outer belt and Jupiter Trojan asteroids, and the remainder is distributed among the inner belt and planet-crossing asteroid populations.

The densities of the largest asteroids as compared with those of the inner planets

The densities of Ceres, Pallas, and Vesta are 2.3, 3.4, and 4.0 grams per cubic centimetre, respectively. These compare with 5.4, 5.2, and 5.5 g/cm^3 for Mercury, Venus, and the Earth, respectively; 3.9 g/cm^3 for Mars; and 3.3 g/cm^3 for the Moon. The density of Ceres is similar to that of a class of meteorites known as carbonaceous chondrites, which contain a larger fraction of volatile material than do ordinary terrestrial rocks and hence have a somewhat lower density (see below *Specific asteroidal source regions for recovered meteorites*). The density of Vesta is similar to those of the rocky planets. Insofar as Ceres, Pallas, and Vesta are typical of asteroids in general, it can be concluded that asteroids are rocky bodies.

Composition. The combination of spectral reflectance measurements (measures of the amount of reflected sunlight at various wavelengths between about 0.3 and 2.6 micrometres) and albedos is used to classify asteroids into various taxonomic groups. If sufficient spectral resolution is available, these measurements also can be used to infer the composition of the surface reflecting the light. This can be done by comparing the asteroid data with that obtained in the laboratory using meteorites or terrestrial rocks or minerals.

By the end of the 1980s, spectral reflectance measurements at visual wavelengths between 0.3 and 1.1 micrometres were available for about 6,000 asteroids, while albedos were determined for roughly 2,000. Both types of data were available for approximately 400 asteroids. Table 24 summarizes the 15 taxonomic classes into which the asteroids can be divided on the basis of such data.

Asteroids of the B, C, F, and G classes have low albedos and spectral reflectances similar to those of carbonaceous chondritic meteorites and assemblages produced by hydrothermal alteration and/or metamorphism of carbonaceous precursor materials. Some C-class asteroids are known to have hydrated minerals on their surfaces, whereas Ceres (a G-class) probably has water present as a layer of permafrost. K- and S-class asteroids have moderate albedos and spectral reflectances similar to the stony-iron meteorites, and they are known to contain significant amounts of silicates and metals, including the minerals olivine and pyroxene on their surfaces. M-class asteroids are moderate-albedo objects, may have significant amounts of nickel-iron metal in their surface material, and exhibit spectral reflectances similar to the nickel-iron meteorites. P-, D-, and T-class asteroids have low albedos and no known meteorite or naturally occurring mineralogical counterparts, but they may contain a large fraction of carbon polymers or organic-rich silicates or both in their surface material. R-class asteroids are very rare; only one (the high-albedo asteroid 349 Dembowska) has been identified with certainty. The surface material of Dembowska has been identified as being most consistent with a pyroxene and olivine-rich composition analogous to the pyroxene-olivine achondrite meteorites. The E-class asteroids have the highest albedos and spectral reflectances that match those of the enstatite achondrite meteorites. V-class asteroids have reflectance properties closely matching those of one particular type of basaltic achondritic meteorite, the eucrite. The match is so good that some believe that the eucrites exhibited in museums are chips from the surface of a V-class asteroid that were knocked off during a major cratering event. The V class is confined to the large asteroid Vesta and a few very small Earth-approaching asteroids. (For additional information about achondrite meteorites, see below *Types of meteorites*.)

Among the larger asteroids (those with diameters greater than 100 kilometres), the C-class asteroids are the most common (accounting for about 65 percent by number) followed, in decreasing order, by the S (15 percent), D (8 percent), P (4 percent), and M (4 percent) classes. The remaining classes constitute less than 4 percent of the population by number. In fact, there are no A-, E-, or Q-class asteroids in this size range, only one member of the R and V classes, and between two and five members of each of the B, F, G, K, and T classes.

The distribution of the taxonomic classes throughout the asteroid belt is highly structured, as can be seen from Figure 60. Some believe this variation with distance from the Sun means that the asteroids formed at or near their present locations and that a detailed comparison of the chemical composition of the asteroids in each region will provide constraints on models for the conditions that may

Table 24: Summary of Asteroid Taxonomic Classes

class	mean albedo	spectral reflectivity (0.3–1.1 micrometres)
C	0.05	neutral, slight absorption at wavelengths of 0.4 micrometre or shorter
D	0.04	very red at wavelengths of 0.7 micrometre or longer
F	0.05	flat
P	0.04	featureless, sloping up into red*
G	0.09	similar to C class but with a deeper absorption at wavelengths of 0.4 micrometre or shorter
K	0.12	similar to S class but with lower slopes
T	0.08	moderately sloped with weak ultraviolet and infrared absorption bands
B	0.14	similar to C class but with shallower slope toward longer wavelengths
M	0.14	featureless, sloping up into red*
Q	0.21	strong absorption features shortward and longward of 0.7 micrometre
S	0.18	very red at wavelengths of less than 0.7 micrometre, typically with an absorption band between 0.9 and 1.0 micrometre
A	0.42	extremely red at wavelengths shorter than 0.7 micrometre and a deep absorption longward of 0.7 micrometre
E	0.44	featureless, sloping up into red*
R	0.35	similar to A class but with slightly weaker absorption bands
V	0.34	very red at wavelengths of less than 0.7 micrometre and a deep absorption and centred near 0.95 micrometre
Other	any	any object not falling into one of the above classes

*Classes E, M, and P are spectrally indistinguishable at these wavelengths and require an independent albedo measurement for unambiguous classification.

have existed within the contracting solar nebula at the time the asteroids were formed.

THE ORIGIN AND EVOLUTION OF THE ASTEROIDS

Available evidence indicates that the asteroids are the remnants of a "stillborn" planet. It is thought that at the time the planets were forming from the low-velocity collisions among asteroid-size planetesimals, one of them grew at a high rate and to a size larger than the others. In the final stages of its formation this planet, Jupiter, gravitationally scattered large planetesimals, some of which may have been as massive as the Earth is today. These planetesimals were eventually either captured by Jupiter or another of the trans-Jovian planets or ejected from the solar system. While they were passing through the inner solar system, however, such large planetesimals strongly perturbed the orbits of the planetesimals in the region of the asteroid belt, raising their mutual velocities to the average five kilometres per second they exhibit today. These high-velocity collisions ended the accretionary collisions by transforming them into catastrophic disruptions. Only objects larger than about 500 kilometres in diameter could have survived collisions with objects of comparable size at collisional velocities of five kilometres per second. Since that time, the asteroids have been collisionally evolving so that, with the exception of the very largest, most present-day asteroids are either remnants or fragments of past collisions.

While breaking down larger asteroids into smaller ones, collisions expose deeper layers of asteroidal material. If asteroids were compositionally homogeneous, this would have no noticeable result. Some of them, however, have become differentiated since their formation. This means that some asteroids, originally formed from so-called primitive material (*i.e.*, material of nonvolatile solar composition), were heated, perhaps by short-lived radionuclides or solar magnetic induction, to the point where their interiors melted and geochemical processes occurred. In certain cases, temperatures became high enough for iron to form. Being denser than other materials, the iron then sank to the centre, forming an iron core and forcing basaltic lavas onto the surface. At least one asteroid with a basaltic surface, Vesta, survives to this day. Other differentiated asteroids were disrupted by collisions that stripped away their crusts and mantles and exposed their iron cores. Still others may have had only their crusts partially stripped away, which exposed surfaces such as those visible today on the A-, E-, and R-class asteroids.

Evidence of differentiation

Collisions were responsible for the formation of the Hirayama families and at least some of the planet-crossing asteroids. A number of the latter enter the Earth's atmosphere, giving rise to sporadic meteors. Larger pieces survive passage through the atmosphere, and some of these end up in museums and laboratories as meteorites (see below *Relationship of meteoroids to asteroids and comets* and *Meteorites, asteroids, and the early solar system*). The very largest produce craters such as Meteor Crater in Arizona in the southwestern United States, and one may even have been responsible for the extinction of the dinosaurs some 65 million years ago. This extinction may have been triggered by an explosion resulting from the impact of an asteroid measuring roughly 10 kilometres in diameter. (Some investigators believe that a cometary body rather than an asteroid may have caused such an explosion.) Fortunately, collisions of this sort are rare. According to current estimates, a few asteroids of 1-kilometre diameter collide with the Earth every 1 million years. (E.F.T.)

Comets

GENERAL CONSIDERATIONS

Basic features. The traditional definition of a comet is a nebulous body with a "hairy" tail that makes a transient appearance in the sky. The word comet comes from the Greek *komētēs,* meaning "hairy one," a description that fits the bright comets noticed by the ancients. Many comets, however, do not develop tails. Moreover, comets are not surrounded by any nebulosity during most of their lifetime. The only permanent feature of a comet is its nucleus, which is a small body that may be seen as a stellar image in large telescopes when tail and nebulosity do not exist, particularly when the comet is still far away from the Sun. Two characteristics differentiate the cometary nucleus from a very small asteroid—namely, its orbit and its chemical nature. A comet's orbit is more eccentric; therefore, its distance to the Sun varies considerably. Its material is more volatile. When far from the Sun, however, a comet remains in its pristine state for eons without losing any volatile components because of the deep cold of space. For this reason, astronomers believe that pristine cometary nuclei may represent the oldest and best-preserved material in the solar system.

Differences between cometary nuclei and asteroids

During a close passage near the Sun, the nucleus of a comet loses water vapour and other more volatile compounds, as well as dust dragged away by the sublimating gases. It is then surrounded by a transient dusty "atmosphere" that is steadily lost to space. This feature is the coma, which gives a comet its nebulous appearance. The nucleus surrounded by the coma makes up the head of the comet. When it is even closer to the Sun, solar radiation usually blows the dust of the coma away from the head and produces a dust tail, which is often rather wide, featureless, and yellowish. The solar wind, on the other hand, drags ionized gas away in a slightly different direction and produces a plasma tail, which is usually narrow with nods and twists and has a bluish appearance.

Designations. In order to classify the chronological appearance of comets, the *Astronomische Nachrichten* ("Astronomical Reports") introduced in 1870 a system of preliminary and final designations that is still used today with only minor modifications. The preliminary designation classifies comets according to their order of discovery, using the year of discovery followed by a lowercase letter in alphabetical order, as in 1987a, 1987b, 1987c, and so forth. Comets are reclassified as soon as possible—usually a few years later—according to their chronological order of passage at perihelion (closest distance to the Sun); a Roman numeral is used in this case, as in 1987 I, 1987 II, 1987 III, and so on. Since the discovery may have taken place at any time either before or after perihelion passage, the two chronologies are not necessarily in the same order, and even the year may change in the final designation. The official designation generally includes the name(s) of its discoverer(s)—with a maximum of three names—preceded by a P/ if the comet is on a periodic orbit. If a person discovers several comets, an Arabic numeral is used after his name, as in 1867 II Tempel 1 and 1873 II Tempel 2. The discoverer's rule has not always been strictly applied: comets P/Halley, P/Lexell, P/Encke, and P/Crommelin have been named after the astronomers who proved their periodic character. Some comets become bright so fast that they are discovered by a large number of persons at almost the same time. They are given an arbitrary impersonal designation such as Brilliant Comet (1882 II), Southern Comet (1947 XII), or Eclipse Comet (1948 XI). Finally, comets may be discovered by an unusual instrument without direct intervention of a specific observer, as in the case of the Earth-orbiting Infrared Astronomical Satellite (IRAS). Its initials are used as if it were a human observer, as in 1983 VII IRAS-Araki-Alcock.

HISTORICAL SURVEY OF COMET OBSERVATIONS AND STUDIES

Early observations. In ancient times, without interference from streetlights or urban pollution, comets could be seen by everyone. Their sudden appearance—their erratic behaviour against the harmonious order of the heavenly motions—was interpreted as an omen of nature that awed people and was used by astrologers to predict flood, famine, pestilence, or the death of kings. The Greek philosopher Aristotle (4th century BC) thought that the heavens were perfect and incorruptible. The very transient nature of comets seemed to imply that they were not part of the heavens but were merely earthly exhalations ignited and transported by heat to the upper atmosphere. Although the Roman philosopher Seneca (1st century AD) had proposed that comets could be heavenly bodies like the planets, Aristotle's ideas prevailed until the 14th century AD. Finally, during the 16th century the Danish no-

The studies of Tycho Brahe

bleman Tycho Brahe established critical proof that comets are heavenly bodies. He compared the lack of diurnal parallax of the comet of 1577 with the well-known parallax of the Moon (the diurnal parallax is the apparent change of position in the sky relative to the distant stars due to the rotation of the Earth). Tycho deduced that the comet was at least four times farther away than the Moon, establishing for the first time that comets were heavenly bodies.

The impact of Newton's work. The German astronomer Johannes Kepler still believed in 1619 that comets travel across the sky in a straight line. It was the English physicist and mathematician Isaac Newton who demonstrated in his *Principia* (1687) that, if heavenly bodies are attracted by a central body (the Sun) in proportion to the inverse square of its distance, they must move along a conic section (circle, ellipse, parabola, or hyperbola). Using the observed positions of the Great Comet of 1680, he identified its orbit as being nearly parabolic.

Newton's friend, the astronomer Edmond Halley, endeavoured to compute the orbits of 24 comets for which he had found accurate enough historical documents. Applying Newton's method, he presupposed a parabola as an approximation for each orbit. Among the 24 parabolas, 3 were identical in size and superimposed in space. The three relevant cometary passages (1531, 1607, and 1682) were separated by two time intervals of 76 and 75 years. Halley concluded that the parabolas were actually the end of an extremely elongated ellipse. Instead of three curves open to infinity, the orbit is closed and brings the same comet periodically back to the Earth. As a consequence, it would return in 1758, he predicted. Observed on Christmas night, 1758, by Johann Georg Palitzsch, a German amateur astronomer, the comet passed at perihelion in March 1759 and at perigee (closest to the Earth) in April 1759. The perihelion date of 1759 had been predicted with an accuracy of one month by Alexis-Claude Clairaut, a French astronomer and physicist. Clairaut's work contributed much to the acceptance of Newton's theory on the Continent. With this, the until-then anonymous comet came to be called Halley's comet (or, in modern nomenclature, Comet P/Halley).

Passages of Comet P/Halley. Since 1759, Comet Halley has reappeared three more times—in 1835, 1910, and 1986. Its trajectory has been computed backward, and all of its 30 previous passages described in historical documents over 22 centuries have been authenticated. Comet Halley's period has irregularly varied between 74.4 years (from 1835 to 1910) and 79.6 years (from AD 451 to 530). These variations, which have been accurately predicted, result from the changing positions of the giant planets, mainly Jupiter and Saturn, whose variable attractions perturb the trajectory of the comet. The space orientation of the orbit has been practically constant, at least for several centuries. Since its returns are not separated by an integer number of years, however, the comet encounters the Earth each time on a different point of its orbit around the Sun; thus, the geometry of each passage is different and its shortest distance to the planet varies considerably. The closest known passage to the Earth, 0.033 AU, occurred on April 9, AD 837.

The 1986 passage of Comet Halley

The perigee distance of most of Comet Halley's historical passages has been between 0.20 and 0.50 AU. The last perigee, on April 11, 1986, took place at 0.42 AU from the Earth (Figure 62). By contrast, the comet passed at only 0.14 AU from the Earth in 1910. Seen from closer range, it was brighter and had a longer tail than on its return in 1986. This is one reason why the 1986 passage proved so disappointing to most lay observers. Yet, a far more important factor had to do with geometry: in the latitudes of the major Western countries, the comet was hidden by the southern horizon during the few weeks in April 1986 when it was at its brightest. Moreover, the night sky of most Western countries is brightly and constantly illuminated by public and private lights. Even in the absence of moonlight, the nighttime sky is pervaded by a milky glare that easily hides the tail of a comet.

Each century, a score of comets brighter than Comet Halley have been discovered. Yet, they appear without warning and will not be seen again. Many are periodic comets like Comet Halley, but their periods are extremely long (millennia or even scores or hundreds of millennia), and they have not left any identifiable trace in prehistory. Bright Comet Bennett 1970 II (Figure 63) will return in 17 centuries, whereas the spectacular Comet West 1976 VI will reappear in about 500,000 years. Among the comets that can easily be seen with the unaided eye, Comet Halley is the only one that returns in a single lifetime. Approximately 100 comets whose periods are between 3 and 200 years are known, however. Unfortunately they are or have become too faint to be readily seen without the aid of telescopes (see below *Periodic comets*).

Figure 62: Comet Halley photographed on March 8 and 9, 1986, by the one-metre Schmidt telescope of the European Southern Observatory at La Silla, Chile.
By courtesy of the European Southern Observatory

Modern cometary research. During the 19th century it was shown that the radiant (*i.e.*, spatial direction) of the spectacular meteor showers of 1866, 1872, and 1885 coincided well with three known cometary orbits that happened by chance to cross the Earth's orbit at the dates of the observed showers. The apparent relationship between

By courtesy of the Department of Astronomy, University of Michigan, Ann Arbor

Figure 63: Comet Bennett, taken at Cerro Tololo Interamerican Observatory, Chile, March 16, 1970.

comets and meteor showers was interpreted by assuming that the cometary nucleus was an aggregate of dust or sand grains without any cohesion. (This conception of the cometary nucleus became known as the "sandbank" model [see below *Cometary models*].) Meteor showers were explained by the spontaneous scattering of the dust grains along a comet's orbit, and the cometary nucleus began to be regarded only as the densest part of a meteor stream. At the end of the 19th and the beginning of the 20th century, spectroscopy revealed that the reflection of sunlight by the dust was not the only source of light in the tail; it showed the discontinuous emission that constitutes the signature of gaseous compounds. More specifically, it revealed the existence in the coma of several radicals—molecular fragments such as cyanogen (CN) and the carbon forms C_2 and C_3, which are chemically unstable in the laboratory because they are very reactive in molecular collisions. Spectroscopy also enabled investigators to detect the existence of a plasma component in the cometary tail by the presence of molecular ions, as, for example, those of carbon monoxide (CO^+), nitrogen (N_2^+), and carbon dioxide (CO_2^+). The radicals and ions are built up by the three light elements carbon (C), nitrogen (N), and oxygen (O). Hydrogen (H) was added when the radical CH was discovered belatedly on spectrograms of Comet Halley taken in 1910. The identification of CH was proposed by the American astronomer Nicholas Bobrovnikoff in 1931 and confirmed in 1938 by Marcel Nicolet of Belgium. In 1941 another Belgian astronomer, Pol Swings, and his coworkers identified three new ions: CH^+, OH^+, and CO_2^+. The emissions of the light elements hydrogen, carbon, oxygen, and sulfur and of carbon monoxide were finally detected when the far ultraviolet spectrum (which is absorbed by the Earth's atmosphere) was explored during the 1970s with the help of rockets and satellites. This included the very large halo (10^7 kilometres) of atomic hydrogen (the Lyman-alpha emission line) first observed in Comets Tago-Sato-Kosaka 1969 IX and Bennett 1970 II.

Spectroscopic discoveries

Although the sandbank model was still seriously considered until the 1960s and '70s by a small minority (most notably the British astronomer Raymond A. Lyttleton), the presence of large amounts of gaseous fragments of volatile molecules in the coma suggested to Bobrovnikoff the release by the nucleus of a bulk of unobserved "parent" molecules such as H_2O, CO_2, and NH_3 (ammonia). In 1948, Swings proposed that these molecules should be present in the nucleus in the solid state as ices.

In a fundamental paper, the American astronomer Fred L. Whipple set forth in 1950 the so-called dirty snowball model, according to which the nucleus is a lumpy piece of icy conglomerate wherein dust is cemented by a large amount of ices—not only water ice but also ices of more volatile molecules. This amount must be substantial enough to sustain the vaporizations for a large number of revolutions. Whipple noted that the nuclei of some comets at least are solid enough to graze the Sun without experiencing total destruction, since they apparently survive unharmed. (Some but not all Sun-grazing nuclei split under solar tidal forces.) Finally, argued Whipple, the asymmetric vaporization of the nuclear ices sunward produces a jet action opposite to the Sun on the solid cometary nucleus. When the nucleus is rotating, the jet action is not exactly radial. This explained the theretofore mysterious nongravitational force identified as acting on cometary orbits. In particular, the orbital period of P/Encke mysteriously decreased by one to three hours per revolution (of 3.3 years), whereas that of P/Halley increased by some three days per revolution (of 76 years). For Whipple, a prograde rotation of the nucleus of P/Encke and a retrograde rotation of that of P/Halley could explain these observations. In each case, a similar amount of some 0.5 to 0.25 percent of the ices had to be lost per revolution to explain the amount of the nongravitational force. Thus, all comets decay in a matter of a few hundred revolutions. This duration is only at most a few centuries for Encke and a few millennia for Halley. At any rate, it is millions of times shorter than the age of the solar system.

Dirty snowball model

The observed comets, however, have obviously survived until now. If they have existed for billions of years, they must have been stored in an extremely cold place far away from the Sun before recently coming into the inner solar system where they could be seen from the Earth. A reply to such a suggestion had already been anticipated in 1932 by the Estonian-born astronomer Ernest J. Öpik, who proposed the possible existence of a large cloud of unobservable comets surrounding the solar system. Nearly 20 years later, the Dutch astronomer Jan Hendrick Oort established the existence of such a cloud of comets by indirect reasoning based on observations. Since the appearance of his theory in 1950, this enormous cloud of comets has come to be called the Oort cloud.

The Oort cloud

Oort showed by statistical arguments that a steady flux of a few "new" comets are observed per year (those that had never been through the solar system before). This flux comes from the fringe of the Oort cloud. He identified it by looking at the distribution of the original values of the total energies of cometary orbits (see discussion of total energy below in *Motion and discovery of comets: Types of orbits*). These energies are in proportion to a^{-1}, with a being the semimajor axis of the cometary orbit. The original value of a refers to the orbit when the comet was still outside of the solar system, as opposed to the osculating orbit, which refers to the arc observed from the Earth after it has been modified by the perturbations of the giant planets. Passages through the solar system produce a rather wide diffusion in orbital energies (in a^{-1}). In 1950 Oort accounted for only 19 accurate original orbits of long-period comets. The fact that 10 of the 19 orbits were concentrated in a very narrow range of a^{-1} established that most of them had never been through this diffusion process due to the planets. The mean value of a for these new comets suggested the distance they were coming from—about 10^5 AU. This distance is also the place where perturbations resulting from the passage of nearby stars begin to be felt. The distance coincidence suggested to Oort that stellar perturbations were the mechanism by which comets were sent into the planetary system.

Subsequent work by the American astronomer Brian G. Marsden and his coworkers confirmed the existence of the Oort cloud. Their list of approximately 90 original orbits crammed within an extremely narrow range of a^{-1} corroborated Oort's initial effort. The mean aphelion distance of this list of new comets implies, however, that the Oort cloud margin is only at some 40,000 to 50,000 AU, which makes the standard mechanism of stellar perturbations much less effective than Oort had believed. Comets must therefore come down from the Oort cloud in several steps, penetrating first into the outer solar system where the perturbations of Uranus and Neptune are weak enough not to remove them from the action of passing stars except after several revolutions.

During the late 1980s astronomers explored new ideas with which to determine how the outer perturbations on the Oort cloud could increase. Dark molecular clouds, for example, may be substituted for stars as major perturbing agents. The hypothesis that there exists some extra undetected matter (like black dwarfs) in the disk of the Milky Way Galaxy also has been used. Then, the total mass distribution in the galactic disk may be large enough to induce tidal forces in the Oort cloud that would change cometary orbits.

MOTION AND DISCOVERY OF COMETS

Types of orbits. In the absence of planetary perturbations and nongravitational forces, a comet will orbit the Sun on a trajectory that is a conic section with the Sun at one focus. The total energy E of the comet, which is a constant of motion, will determine whether the orbit is an ellipse, a parabola, or a hyperbola. The total energy E is the sum of the kinetic energy of the comet and of its gravitational potential energy in the gravitational field of the Sun. Per unit mass, it is given by $E = \frac{1}{2}v^2 - GMr^{-1}$, where v is the comet's velocity and r its distance to the Sun, with M denoting the mass of the Sun and G the gravitational constant. If E is negative, the comet is bound to the Sun and moves in an ellipse. If E is positive, the comet is unbound and moves in a hyperbola. If $E = 0$, the comet is unbound and moves in a parabola.

Total energy of a comet

In polar coordinates written in the plane of the orbit, the general equation for a conic section is

$$r = q(1 + e)(1 + e \cos \theta)^{-1},$$

where r is the distance from the comet to the Sun, q the perihelion distance, e the eccentricity of the orbit, and θ an angle measured from perihelion. When $0 \leq e < 1$, $E < 0$ and the orbit is an ellipse (the case $e = 0$ is a circle, which constitutes a particular ellipse); when $e = 1$, $E = 0$ and the orbit is a parabola; and when $e > 1$, $E > 0$ and the orbit is a hyperbola.

Orbital elements

In space a comet's orbit is completely specified by six quantities called its orbital elements. Among these are three angles that define the spatial orientation of the orbit: i, the inclination of the orbital plane to the plane of the ecliptic; Ω, the longitude of the ascending node measured eastward from the vernal equinox; and ω, the angular distance of perihelion from the ascending node (also called the argument of perihelion). The three most frequently used orbital elements within the plane of the orbit are q, the perihelion distance in astronomical units; e, the eccentricity; and T, the epoch of perihelion passage.

Identifying comets and determining their orbits. Up to the beginning of the 19th century, comets were discovered exclusively by visual means. Many discoveries are still made visually with moderate-size telescopes by amateur astronomers. Although comets can be present in any region of the sky, they are often discovered near the western horizon after sunset or near the eastern horizon before sunrise, since they are brightest when closest to the Sun. Because of the Earth's rotation and direction of motion in its orbit, discoveries before sunrise are more likely, as confirmed by discovery statistics. At discovery a comet may still be faint enough not to have developed a tail; therefore, it may look like any nebulous object—*e.g.*, an emission nebula, a globular star cluster, or a galaxy. The famous 18th-century French comet hunter Charles Messier (nicknamed "the ferret of comets" by Louis XV for his discovery of 21 comets) compiled his well-known catalog of "nebulous objects" so that such objects would not be mistaken for comets. The final criterion remains the apparent displacement of the comet after a few hours or a few days with respect to the distant stars; by contrast, the nebulous objects of Messier's catalog do not move. After such a displacement has been undisputably observed, any amateur wishing to have the comet named for himself must report his claim to the nearest observatory as soon as possible.

Photographic discovery of comets

Most comets are and remain extremely faint. Today, a larger and larger proportion of comet discoveries are thus made fortuitously from high-resolution photographs, as, for instance, those taken during sky surveys by professional astronomers engaged in other projects.

The faintest recorded comets are approaching the limit of detection of large telescopes (those that are two metres or more in diameter). That is to say, they are of the 22nd–23rd magnitude, or 10^6 to 10^7 times fainter than the limit of the naked eye. Several successive photographic observations of these faint moving objects are necessary to ensure identification and simultaneous calculation of a preliminary orbit. In order to determine a preliminary orbit as quickly as possible, the eccentricity $e = 1$ is assumed since some 90 percent of the observed eccentricities are close to one, and a parabolic motion is computed. This is generally sufficient to ensure against "losing" the comet in the sky.

The best conic section representing the path of the comet at a given instant is known as the osculating orbit. It is tangent to the true path of the chosen instant, and the velocity at that point is the same as the true instantaneous velocity of the comet. Nowadays, high-speed computers make it possible to produce a final ephemeris (table of positions) that is not only based on the definitive orbit but also includes the gravitational forces of the Sun and of all significant planets that constantly change the osculating orbit. In spite of this fact, the deviation between the observed and the predicted positions usually grows (imperceptibly) with the square of time. This is the signature of a "neglected" acceleration, which comes from a nongravitational force (see above). Formulas representing the smooth variation of the nongravitational force with heliocentric distance are now included for many orbits. The most successful formula assumes that water ice prevails and controls the vaporization of the nucleus.

COMETARY STATISTICS

The *Catalog of Cometary Orbits,* compiled by Marsden, remains the standard reference for orbital statistics. Its 1989 edition lists 1,292 computed orbits from 239 BC to AD 1989; only 91 of them were computed using the rare accurate historical data from before the 17th century. More than 1,200 are therefore derived from cometary passages during the last three centuries. The 1,292 cometary apparitions of Marsden's catalog involve only 810 individual comets; the remainder represents the repeated returns of periodic comets.

Periodic comets. The periodic comets are usually divided into short-period comets (those with periods of less than 200 years) and long-period comets (those with periods of more than 200 years). Of the 155 short-period comets, 93 have been observed at two or more perihelion passages. In 1989, four of these comets had been definitely lost, and three more were probably lost, presumably because of their decay in the solar heat. Some authors have found it advantageous to change the definition of short-period comets by diminishing their longest-period cutoff to 20 years. This leaves 135 short-period comets (new style) in the *Catalog;* the 20 others having periods between 20 and 200 years are called intermediate-period comets. These two new classes are separated by a small period gap. The average short-period comet has a seven-year period, a perihelion distance of 1.5 AU, and a small inclination (13°) on the ecliptic. All short-period comets (new style) revolve in the direct (prograde) sense around the Sun, just as the planets do. The intermediate-period comets have on average a larger inclination of the ecliptic, and five of them turn around the Sun in a retrograde direction. The most famous of the latter is P/Halley (30 appearances); the others are P/Tempel-Tuttle (4 appearances), P/Pons-Gambart, P/Hartley-IRAS, and P/Swift-Tuttle (the last three with only 1 appearance each). Eleven of the 20 intermediate-period comets have been observed during a single appearance.

Short- and intermediate-period comets

Long-period comets

The comets with long-period orbits are distributed at random in all directions of the sky, and roughly half of them turn in the retrograde direction. Of the 655 comets of long period contained in the *Catalog,* 192 have osculating elliptic orbits, and 122 have osculating orbits that are very slightly hyperbolic. Finally, 341 are listed as having parabolic orbits, but this is rather fallacious because either it has not been possible to detect unequivocal deviations from a parabola on the (sometimes very short) arc along which the comets have been observed or, more simply, the final calculations have never been made. The parabola is always assumed first in the preliminary computation as it is easier to deal with. If the osculating orbit is computed backward to when the comet was still far beyond the orbit of Neptune and if the orbit is then referred to the centre of mass of the solar system, the original orbits almost always prove to be elliptic. (The centre of mass of the solar system is different from the centre of the Sun primarily because of the position of massive Jupiter.) Twenty-two original orbits remain (nominally) slightly hyperbolic beyond the orbit of Neptune, but 19 remain not significantly different from a parabola. Even the three that are significantly different near 50 AU are likely to become elliptic when they are 50,000 or 100,000 AU from the Sun. The reason is that, though the mass of the Oort cloud remains uncertain, it should be added to the mass of the inner solar system to compute the orbits. The smallest possible mass of the Oort cloud is likely to transform the orbits into ellipses. It is thus reasonable to believe that all observed comets were initially in elliptic orbits bound to the solar system. Accordingly, all parabolic and nearly parabolic comets are thought to be comets of very long period.

The future orbit of a long-period comet is obtained when the osculating orbit is computed forward to when the comet will be leaving the planetary system (beyond the orbit of Neptune) and is referred to the centre of mass of

the solar system. Because of the planetary perturbations, slightly more than half of the future orbits become strongly elliptic, whereas slightly less than half become strongly hyperbolic. Roughly half of the long-period comets are thus "captured" by the solar system on more strongly bound orbits; the other half are permanently ejected out of the system.

"New" comets

Among the very-long-period comets, there is a particular class that Oort showed as having never passed through the planetary system before (see above), notwithstanding the fact that their original orbits were elliptic, which implies repeated passages. This paradox vanishes when it is understood that their perihelia were outside of the planetary system before their first appearance but that their orbits have been perturbed near aphelia (either by stellar or dark interstellar-cloud passages or by galactic tides) in such a way that their perihelia were lowered into the planetary system. The first passage of a "new" comet is usually brighter than an average passage (a large fraction of the famous bright historical comets were such new comets). This is possibly explained by the presence of more volatile gases and of a larger component of very fine dust. The most volatile gases may have disappeared during subsequent passages, and the finest dust may have agglomerated into larger dust grains that reflect less light for the same production rate. About 90 comets have been identified as new in long-period orbits. If the same proportion exists in the poorly computed parabolic orbits, the total must be close to 170 new comets in Marsden's catalog, but 80 of them have not been identified.

Groups of comets and other unusual cometary objects. Some comets travel in strikingly similar orbits, only the time of perihelion passages being appreciably different. Members of such a group of comets are thought to be fragments from a larger comet that was tidally disrupted earlier by the Sun or in some cases by the differential jet action of nongravitational forces on a fragile nucleus. Many such breakups have been observed historically. Slight differences in the resultant velocities—though they occur very gently—are sufficient to cause cometary fragments to separate along orbits close to but distinct from each other, particularly as far as their total energy is concerned. A very slight variation in a^{-1} introduces an orbital period that may vary by several years, and when the cometary fragments return they will go through perihelion at widely separated epochs. The best-known example is the famous group of "Sun-grazing" comets (also called the Kreutz group), which has 12 definite members (plus one probable) with perihelion distances between 0.002 and 0.009 AU (less than half a solar radius). Their periods are scattered from 400 to 2,000 years, and their last passages occurred between 1880 and 1970. The most famous fragment of the group is Comet Ikeya-Seki 1965 VIII.

Comet P/Schwassmann-Wachmann 1, which has a period of 15 years, is in a quasi-circular and somewhat unstable orbit between Jupiter and Saturn, with a perihelion q that equals 5.45 AU and an aphelion of 6.73 AU. It can be observed every year for several months when opposite to the Sun in the sky. Without any visible tail, it has irregular outbursts that make its coma grow in size for a few weeks and become up to 1,000 times as bright as normal.

Unusual comets

Another unusual object is the so-called asteroid 2060 Chiron, which has a similar orbit between Saturn and Uranus. Though first classified as an asteroid, its icy nucleus of some 300 kilometres suggests that it is a giant comet provisionally parked on a quasi-circular but unstable orbit. Indeed, Chiron develops weak, sporadic outbursts, and in 1989 a transient nebulosity surrounding it (a "coma") was reported for the first time. Within a few thousand years, Chiron might be perturbed enough by Saturn to come closer to the Sun and become a spectacular comet.

For faraway objects that contain volatile ices, the distinction between asteroids and comets becomes a matter of semantics because many orbits are unstable; an asteroid that comes closer to the Sun than usual may become a comet by producing a transient atmosphere that gives it a fuzzy appearance and that may develop into a tail. Some objects have been reclassified as a result of such occurrences. For example, asteroid 1990 UL3, which crosses the orbit of Jupiter, was reclassified as Comet P/Shoemaker-Levy 2 late in 1990. Conversely, it is suspected that some of the Earth-approaching asteroids (Amors, Apollos, and Atens) could be the extinct nuclei of comets that have now lost most of their volatile ices (see above *Orbits of asteroids; Near-Earth asteroids*).

Two bright comets, Morehouse 1908 III and Humason 1962 VIII, exhibited a peculiar tail spectrum in which the ion CO^+ prevailed in a spectacular way, possibly because of an anomalous abundance of a parent molecule (carbon monoxide, carbon dioxide, or possibly formaldehyde [CH_2O]) vaporizing from the nucleus. Finally, Comet Halley is the brightest and therefore the most famous of all short- and intermediate-period comets as the only one that returns in a single lifetime and can be seen with the naked eye.

THE NATURE OF COMETS

The nucleus. As previously noted, the traditional picture of a comet with a hazy head and a spectacular tail applies only to a transient phenomenon produced by the decay in the solar heat of a tiny object known as the cometary nucleus. In the largest telescopes, the nucleus is never more than a bright point of light at the centre of the cometary head. At substantial distances from the Sun, the comet seems to be reduced to its starlike nucleus. The nucleus is the essential part of a comet because it is the only permanent feature that survives during the entire lifetime of the comet. In particular, it is the source of the gases and dust that are released to build up the coma and tail when a comet approaches the Sun. The coma and tail are enormous: typically the coma measures 100,000 kilometres or more in diameter, and the tail may extend about 100,000,000 kilometres in length. They scatter and continuously dissipate into space but are steadily rebuilt by the decay of the nucleus, whose size is usually in the range of 10 kilometres.

The evidence on the nature of the cometary nucleus remained completely circumstantial until March 1986, when the first close-up photographs of the nucleus of Comet Halley were taken during a flyby by the *Giotto* spacecraft of the European Space Agency (Figure 64). Whipple's basic idea that the cometary nucleus was a monolithic piece of icy conglomerate (see above) had been already

Figure 64: Composite image of the nucleus of Comet Halley produced from 68 original photographs taken by the Halley Multicolour Camera on board the *Giotto* spacecraft on March 13 and 14, 1986.

well supported by indirect deductions in the 1960s and '70s and had become the dominant though not universal view. The final proof of the existence of such a "dirty snowball," however, was provided by the photographs of Comet Halley's nucleus.

If there was any surprise, it was not over its irregular shape (variously described as a potato or a peanut), which had been expected for a body with such small gravity (10^{-4} *g*, where *g* is the gravity of the Earth). Rather, it was over the very black colour of the nucleus, which suggests that the snows or ices are indeed mixed together with a large amount of sootlike materials (*i.e.,* carbon and tar in fine dust form). The very low geometric albedo (2 to 4 percent) of the cometary nucleus puts it among the darkest objects of the solar system. Its size is thus somewhat larger than anticipated: the roughly elongated body measures 15 by 8 kilometres and has a total volume of some 500 cubic kilometres. Its mass is rather uncertain, estimated in the vicinity of 10^{17} grams, and its bulk density is very small, ranging anywhere from 0.1 to 0.8 gram per cubic centimetre. The infrared spectrometer on board the Soviet Vega 2 spacecraft estimated a surface temperature of 300 to 400 K for the inactive "crust" that seems to cover 90 percent of the nucleus. Whether this crust is only a warmer layer of outgassed dust or whether the dust particles are really fused together by vacuum welding under contact is still open to speculation.

The 10 percent of the surface of Halley's nucleus that shows signs of activity seems to correspond to two large and a few smaller circular features resembling volcanic vents. Large sunward jets of dust originate from the vents; they are clearly dragged away by the gases vaporizing from the nucleus. This vaporization has to be a sublimation of the ices that cools them down to no more than 200 K in the open vents. The chemical composition of the vaporizing gases, as expected, is dominated by water vapour (about 80 percent of the total production rate). The next most abundant volatile (close to 10 percent) appears to be carbon monoxide (CO), though it could come from the dissociation of another parent molecule (*e.g.,* carbon dioxide [CO_2] or formaldehyde [CH_2O]). Following CO in abundance is CO_2 (close to 4 percent). Methane (CH_4) and ammonia (NH_3), on the other hand, seem to be close to the 0.5 to 1 percent level, and the percentage of carbon disulfide (CS_2) is even lower; at that level, there also must be unsaturated hydrocarbons and amino compounds responsible for the molecular fragments observed in the coma. This is not identical to—though definitely reminiscent of—the composition of the volcanic gases on the Earth, which also are dominated by water vapour, but their CO_2:CO, CO_2:CH_4, and SO_2:S_2 ratios are all larger than in Comet Halley, meaning that the volcanic gases are more oxidized. The major difference may stem from the different temperature involved—often near 1,300 K in terrestrial volcanoes, as opposed to 200 K for cometary vaporizations. This may make the terrestrial gases closer to thermodynamic equilibrium. The dust-to-gas mass ratio is uncertain but is possibly in the vicinity of 0.4 to 1.1.

Chemical composition of Comet Halley's nucleus

The dust grains are predominantly silicates. Mass spectrometric analysis by the *Giotto* spacecraft revealed that they contain as much as 20–30 percent carbon, which explains why they are so black. There also are grains composed almost entirely of organic material (molecules made of atoms of hydrogen, carbon, nitrogen, and oxygen).

There is some uncertainty concerning the rotation of Halley's nucleus. Two different rotation rates of 2.2 days and 7.3 days have been deduced by different methods. Both may exist, one of them involving a tumbling motion, or nutation, that results from the irregular shape of the nucleus, which has two quite different moments of inertia along perpendicular axes.

Scientific knowledge of the internal structure of the cometary nucleus was not enhanced by the flyby of Comet Halley, and so it rests on weak circumstantial evidence from the study of other comets. Earlier investigations had established that the outer layers of old comets were processed by solar heat. These layers must have lost most of their volatiles and developed a kind of outgassed crust, which probably measures a few metres in thickness. Inside the crust there is thought to exist an internal structure that is radially the same at any depth. Arguments supporting this view are based on the fact that cometary comas and tails do not become essentially different when comets decay. Since they lose more and more of their outer layers, however, the observed phenomena come from material from increasingly greater depths. These arguments are specifically concerned with the dust-to-gas mass ratio, the atomic and molecular spectra, the splitting rate, and the vaporization pattern during fragmentation.

Before the *Giotto* flyby of Comet Halley, other cometary nuclei had never been resolved optically. For this reason, their albedos had to be assumed first in order to compute their sizes. Techniques proposed to deduce the albedo yielded only that of the dusty nuclear region made artificially brighter by light scattering in the dust. In 1986 the albedo of Comet Halley's nucleus was found to be very low (A = 2 to 4 percent). If this value is typical for other comets, then 11 of 18 short-period comets studied would be between 6 and 10 kilometres in diameter; only 7 of them would be somewhat outside these limits. Comet Schwassmann-Wachmann 1 (see above) would be a giant with a diameter of 96 kilometres; 10 long-period comets would all have diameters close to 16 kilometres (within 10 percent). Since short-period comets have remained much longer in the solar system than comets having very long periods, the smaller size of the short-period comets might result from the steady fragmentation of the nucleus by splitting. Yet, the albedo may also diminish with aging. At the beginning, if the albedo were close to that of slightly less dirty snow (A = 10 percent), the nuclear diameter of long-period comets would come very close to that of the largest of the short-period comets. The diameters of new comets also have been shown to be rather constant and most likely measure close to 10 kilometres. Of course, these are mean "effective" diameters of unseen bodies that are all likely to be very irregular.

Variation in size of nuclei

The region around the nucleus, up to 10 or 20 times its diameter, contains an amount of dust large enough to be partially and irregularly opaque or at least optically thick. It scatters substantially more solar light than is reflected by the black nucleus. Dust jets develop mainly sunward, activated by the solar heat on the sunlit side of the nucleus. They act as a fountain that displaces somewhat the centre of light from the centre of mass of the nucleus. This region also is likely to contain large clusters of grains that have not yet completely decayed into finer dust; the grains are cemented together by ice.

The gaseous coma. The coma, which produces the nebulous appearance of the cometary head, is a short-lived, rarefied, and dusty atmosphere escaping from the nucleus. It is seen as a spherical volume having a diameter of 10^5 to 10^6 kilometres, centred on the nucleus. The coma gases expand at a velocity of about 0.6 kilometre per second. This velocity can be measured from the motion of expanding "halos" triggered by outbursts in the nucleus, from the speed required to produce the Greenstein effect (see below), and from the fluid dynamics required to drag dust particles away at those places where they are observed in the dust tails. This expansion velocity, *v*, varies somewhat with heliocentric distance *r*: $v = 0.58r^{-0.5}$ (in kilometres per second, when *r* is in astronomical units). The light of the spherical coma comes mainly from molecular fragments that have been produced by the dissociation of unobserved "parent molecules" in a zone on the order of 10^4 kilometres around the nucleus. This also is the approximate size of the zone where molecular collisions continue to occur; beyond that zone, the gas becomes too rarefied for such interaction to occur. The zone simply expands radially without molecular collisions into the vacuum of space. The parent molecules (*e.g.,* those of water vapour, carbon dioxide, and hydrogen cyanide [HCN]) are generally not observed because they do not fluoresce in visible light. So far, only a few have been observed at millimetre or centimetre wavelengths by radio telescopes; many more are needed if they are to be regarded as the source of the various radicals and ions that have been detected (Table 25).

Molecular collision zone

If the mixture of original parent molecules has been frozen out of thermodynamic equilibrium in the nuclear

ices, many chemical reactions can still take place in the molecular collision zone. At the usually cold temperature of vaporization, the kinetics of fast ion-molecular reactions would prevail. The reactions might reshuffle the original molecules present in the nucleus into new parent species, which would be the ones subsequently photodissociated into observed fragments by solar light. (This complex situation is still far from being completely understood.) In turn, the observed fragments, after having absorbed and reemitted photons from the solar light several times, would photodissociate or photoionize, which make them disappear from sight at the fuzzy limit of the light-emitting coma (typically $2–5 \times 10^5$ kilometres). A composite list of all observed species in cometary comas and tails is given in Table 25. It is based mainly on observations of the bright comets of the 1960s, '70s, and '80s, including spacecraft results from Comet Halley.

The organic radicals given in Table 25 were seen in cometary heads as visual or ultraviolet emission lines or bands. The exceptions were water vapour, along with hydrogen cyanide and methyl cyanide (CH_3CN); these species, which could be called parent molecules, were observed as pure rotation lines at radio frequencies. The metals—except for sodium (Na), which is observed in many comets—were seen as visual lines in Sun-grazing comets alone. They are assumed to result from the vaporization of dust grains by solar heat. Sodium is a volatile metal that is not unlikely to vaporize easily from dust grains at large distances from the Sun (more than 1 AU). The ions were seen in the visual or ultraviolet emission lines or bands at the onset of the plasma tail or detected by spacecraft. The silicate signature was found in infrared emission bands at the onset of dust tails. The occurrence of the silicate elements, as well as the presence of a rather large amount of organic compounds, was confirmed by the mass spectrometric analysis of dust grains during the *Giotto* flyby of Comet Halley.

An extremely weak coma appeared in 1984 when Comet Halley still was 6 AU from the Sun. In February 1991, the Belgian astronomers Olivier Hainaut and Alain Smette detected a giant outburst from Comet Halley, which was already at a distance of 14.5 AU from the Sun and had the form of a fanlike structure in the direction of the Sun; this is the best case study to date. Rarely have comas been detected beyond 3 or 4 AU, where they are still quite small; they grow to a maximum near 1.5 AU and seem to contract as they approach closer to the Sun. This effect comes from the more rapid decay in solar light (by photoionization or photodissociation) of the visible radicals that emit the coma light. The discrete emission of light by cometary atoms, radicals, or ions is due to the selective absorption of sunlight followed by its reemission either at the same wavelength (resonance) or at a different wavelength (fluorescence). In 1941, Pol Swings explained the peculiar appearance of some of the molecular bands in comets by the irregular spectral distribution of the exciting solar radiation owing to the presence of Fraunhofer lines (dark, or absorption, lines) in this radiation. The temporal variations that occur in the molecular bands as a comet approaches the Sun were explained quantitatively by the variable shift in the apparent wavelengths of the solar Fraunhofer lines due to the variable radial velocity of the comet. This is the so-called Swings effect. Later, the American astronomer Jesse Greenstein explained, by a differential Swings effect, the observed differences in the molecular bands in front of and behind the nucleus: the radial expansion velocity of the coma introduces a different shift forward and backward. This differential Swings effect is often referred to as the Greenstein effect (Figure 65).

Swings effect

Exceptions to the resonance-fluorescence mechanism are known and are exemplified by the case of the emission

Table 25: Observed Chemical Species in Comets

Organic	C, C_2, C_3, CH, CN, CO, CO_2, CS, HCN, CH_3CN, HCO, H_2CO
Inorganic	H, NH, NH_2, O, OH, H_2O, S, S_2, NH_3, NH_4
Metals	Na, K, Ca, V, Mn, Fe, Co, Ni, Cu
Ions	C^+, CH^+, CO^+, CO_2^+, N_2^+, O^+, OH^+, H_2O^+, H_3O^+, S^+, S_2^+, H_2S^+, CS_2^+
Dust	silicates, organic compounds

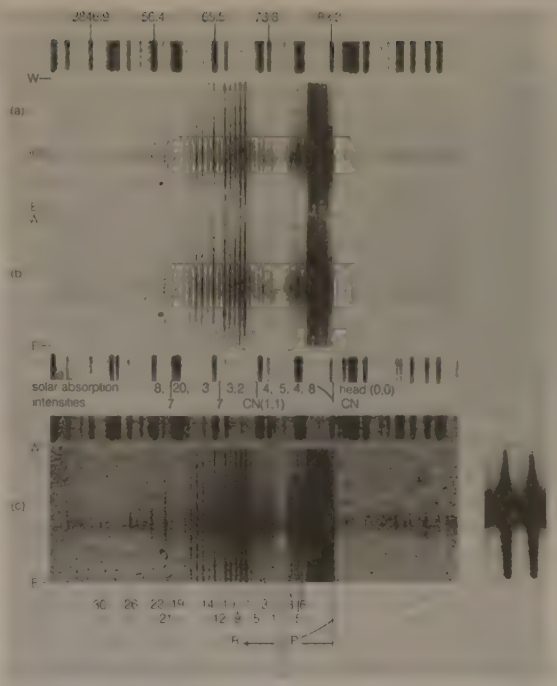


Figure 65: High-dispersion spectrum of the cyanogen (CN) bands near 3880 angstroms in Comet Mrkos, showing band irregularities (Swings and Greenstein effects; see text).
By courtesy of J.L. Greenstein, Palomar Observatory, California Institute of Technology

of the "forbidden" red doublet of atomic oxygen at wavelengths of 6300 and 6364 angstroms. Such an emission cannot be excited by direct absorption of sunlight but is produced directly by the photodissociation of H_2O into $H_2 + O$ (in the ^1D state) and, in an accessorial manner, of CO_2 into $CO + O$ (in the ^1D state). The ^1D state is an excited state of the oxygen atom that decays spontaneously into the ground (lowest energy) state by emitting the forbidden red doublet, provided that it had not been quenched earlier by molecular collisions.

The large atomic hydrogen halo detected up to 10^7 kilometres from the nucleus is simply a large coma visible in ultraviolet (Lyman-alpha line). It is two orders of magnitude larger than the comas that can be seen in visible light only because the hydrogen atoms, being lighter, move radially away 10 times faster and are ionized 10 times more slowly than the other radicals.

Halo of atomic hydrogen

Cometary tails. The tails of comets are generally directed away from the Sun. They rarely appear beyond 1.5 or 2 AU but develop rapidly with shorter heliocentric distance. The onset of the tail near the nucleus is first directed toward the Sun and shows jets curving backward like a fountain, as if they were pushed by a force emanating from the Sun. The German astronomer Friedrich Wilhelm Bessel began to study this phenomenon in 1836, and Fyodor A. Bredikhin of Russia developed, in 1903, tail kinematics based on precisely such a repulsive force that varies as the inverse square of the distance to the Sun. Bredikhin introduced a scheme for classifying cometary tails into three types, depending on whether the repulsive force was more than 100 times the gravity of the Sun (Type I) or less than one solar gravity (Types II and III). Subsequent research showed that Type-I tails are plasma tails (containing observed molecular ions as well as electrons not visible from ground-based observatories), and Types II and III are dust tails, the differences between them being attributable to a minor difference in the size distribution of the dust grains. As a result of these findings, the traditional classification formulated by Bredikhin is no longer considered viable and is seldom used. Most comets (but not all) simultaneously show both types of tail: a bluish plasma tail, straight and narrow with twists and nods, and a yellowish dust tail, wide and curved, which is often featureless (see Figure 66).

The plasma tail has its onset in a region extremely close

Plasma tail

to the nucleus. The ion source lies deep in the collision zone (typically 1,000 kilometres). It is likely that charge-exchange reactions compete with the photoionization of parent molecules, but the mechanism that produces ions is not yet quantitatively understood. In 1951 the German astronomer Ludwig Biermann predicted the existence of the solar wind (see above) in order to account for the rapid accelerations observed in plasma tails as well as their aberration (*i.e.,* deviation from the direction directly opposite the Sun). The cometary plasma is blown away by the magnetic field of the solar wind until it reaches its own velocity—nearly 400 kilometres per second. This action explains the origin of the large forces postulated by the Bessel-Bredikhin theory. Spectacular changes observed in the plasma tail, such as its sudden total disconnection, have been explained by discontinuous changes in the solar wind flow (*e.g.,* the passage of magnetic sector boundaries).

In 1957 the Swedish physicist Hannes Alfven predicted the draping of the magnetic lines of the solar wind around the cometary ionosphere. This phenomenon was detected by the International Cometary Explorer spacecraft, launched by the U.S. National Aeronautics and Space Administration (NASA), when it passed through the onset of the plasma tail of Comet P/Giacobini-Zinner on Sept. 11, 1985. Two magnetic lobes separated by a current-carrying neutral sheet were observed as expected. A related feature known as the ionopause was detected by the *Giotto* space probe during its flyby of Comet Halley in 1986. The ionopause is a cavity without a magnetic field that contains only cometary ions and is separated from the solar wind by a sharp discontinuity. Halley's ionopause lies about 4,000 to 5,000 kilometres from the nucleus of the comet. An analysis of all the encounter data indicates that a complete understanding of cometary interaction with the solar wind has not yet been achieved. It is well understood, however, that the neutral coma remains practically spherical. The solar wind is so rarefied that there are no direct collisions of its particles with the neutral particles of the coma, and, as these particles are electrically neutral, they do not "feel" the magnetic field.

Dust tail

The source of the dust tail is the dust dragged away by the vaporizing gases that emanate from the active zones of the nucleus, presumably from vents like those observed on Comet Halley's nucleus (Figure 64). The dust jets are first directed sunward but are progressively pushed back by the radiation pressure of sunlight. The repulsive acceleration of a particle varies as $(sd)^{-1}$ (with linear size s and density d). For a given density, it thus varies as s^{-1}, separating widely the particles of different sizes in different parts of the tail. Studying the dust tail isophotes of varying brightnesses therefore yields the dust grain distribution. This distribution may peak for very fine particles near 0.5 micrometre (μm), assuming a density of two, as in the case of Comet Bennett; however, it falls off with s^{-n} (with n ranging from three to five) for larger particles. This mechanism neglects particles much smaller than the mean wavelength of sunlight. Because such particles do not reflect light, they do not feel its radiation pressure. (They are not detected from ground-based observations anyway.)

One of the major results of the *Giotto* flyby of Halley's nucleus was the detection of abundant particles much smaller than the wavelength of light, indicating that the size distribution does not peak near 0.5 μm but seems rather to grow indefinitely with a slope close to a^{-2} for finer and finer particles down to possibly 0.05 μm (10^{-17} gram). The dust composition analyzers on board the *Giotto* and Vega spacecraft revealed the presence of at least three broad classes of grains. Class 1 contains the light elements hydrogen, carbon, nitrogen, and oxygen only (in the form of either ices or polymers of organic compounds). The particles of class 2 are analogous to the meteorites known as CI carbonaceous chondrites but are possibly slightly enriched in carbon and sulfur. Class 3 particles are even more enriched in carbon, nitrogen, and sulfur; they could be regarded as carbonaceous silicate cores (like those of class 2) covered by a mantle of organic material (similar to that of class 1) that has been radiation-processed. Most of the encounter data were excellent for elemental analyses but poor for determining molecular composition, because most molecules were destroyed by impact at high encounter velocity. Hence, there still remains much ambiguity regarding the chemical nature of the organic fraction present in the grains.

Meteors are extraterrestrial particles of sand-grain or small-pebble size that become luminous upon entering the upper atmosphere at very high speeds. Meteor streams have well-defined orbits in space. More than a dozen of these orbits have practically the same orbital elements as the orbits of the identical number of short-period comets (see below *Orbits of meteoroids and meteor showers*). Fine cometary dust consists primarily of micrometre- or sub-micrometre-size particles that are much too small to become visible meteors (they are more like cigarette smoke than dust). Moreover, they are scattered in the cometary tail at great distance from the comet orbit. The size distribution of cometary dust grains, however, covers many orders of magnitude; a small fraction of them may reach 0.1 millimetre to even a few centimetres. Because of their large size, these dust grains are almost not accelerated by the radiation pressure of sunlight. They remain in the plane of the cometary orbit and in the immediate vicinity of the orbit itself, even though they separate steadily from the nucleus. They sometimes become visible as an anti-tail—*i.e.,* as a bright spike extending from the coma

Anti-tail

By courtesy of Mount Wilson Observatory

Figure 66: Comet Mrkos, photographed Aug. 22, 24, 26, 27, 1957. The straight tail with prominent streamers and irregularities, best seen at upper left in all photographs, is formed of ionized molecules; the more uniform, curved tail is composed primarily of small solid particles.

Figure 67: Comet Arend-Roland photographed on April 25, 1957. The prominent anti-tail extending from the coma appears to precede the comet, though it actually trails from behind.

By courtesy of Lick Observatory, University of California

sunward in a direction opposite to the tail (Figure 67). This phenomenon occurs as a matter of geometry: it takes place for only a few days when the Earth crosses the plane of the cometary orbit. At such a time, this plane is viewed through the edge, and all large grains are seen accumulated along a line. The same grains scatter farther and farther away from the nucleus until some are along the entire cometary orbit. When the Earth's orbit intersects such an orbit (an event that occurs year after year at the same calendar date), these large grains produce meteor showers.

Extremely fine cometary grains also may penetrate the Earth's atmosphere, but they can be slowed down gently without burning up. Some have been collected by NASA's U-2 aircraft at very high altitudes. Grains of this kind are known as Brownlee particles and are believed to be of cometary origin (Figure 68). Their composition is chondritic, though they show somewhat more carbon and sulfur than the CI carbonaceous chondrites, and their structure is fluffy with many pores. Similar grains were found in space during the space probe exploration of Comet Halley.

By courtesy of D. Brownlee, University of Washington, photograph, M. Wheelock

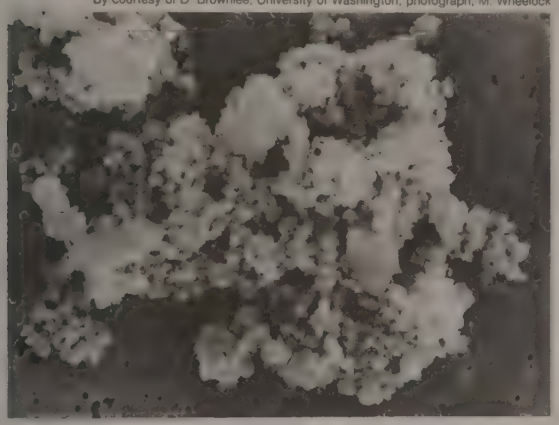


Figure 68: Electron micrograph of chondritic interplanetary dust particle (18.3 micrometres in width) of possible cometary origin. The particle was collected in the Earth's atmosphere by a NASA U-2 research plane.

COMETARY MODELS

As previously noted, the sandbank model of the cometary nucleus fell into disregard by the late 1950s and early 1960s and was supplanted by the dirty snowball (or icy conglomerate) concept. Much circumstantial evidence supported the latter, but confirmation was lacking until 1986, when the *Giotto* spacecraft returned detailed, close-up photographs of Comet Halley's nucleus. Yet, while these photographs corroborated the general idea of the model, they revealed that "dirty snowball" was in fact a misnomer because snow (even when dirty) is suggestive of something white or at least gray in colour. In actuality, the cometary nucleus proved to be pitch black owing to the large amount of very fine, black sootlike particles intermixed with the volatile ices (see above).

Many variations of the icy conglomerate model have been proposed since the early 1980s, as, for example, the fractal model, rubble-pile model, and icy-glue model. These names, however, suggest only slightly different types of accretion of primordial particles; they all share common features—namely, irregular shape, heterogeneous mixture, and very low density because of cavities and pores. The existence of a crust or dust mantle of a different nature had already been proposed before the 1986 spacecraft encounter with Comet Halley for two reasons. First, cosmic-ray processing of the outer layers had been described by Leonid M. Shul'man of the Soviet Union (1972) and later advocated by Fred Whipple and Bertram Donn of the United States, while the outgassing of the outer layers by solar heat had also been assumed since the proposal of Whipple's model (1950). Second, detailed models of the formation and disruption of such mantles due to solar-radiation processing of the upper layers had been studied by Devamitta Asoka Mendis of the United States (1979) and M. Horanyi of Hungary (1984).

Elemental abundances of the cometary nucleus

An average heuristic model for the elemental abundances of the cometary nucleus was developed by the American astronomer Armand H. Delsemme in 1982. Delsemme computed the H:C:N:O:S ratios from ultraviolet and visual observations of atomic and molecular species in bright comets detected during the 1970s and deduced the abundances of metals from the chondritic composition of cometary dust. In this model, hydrogen was depleted by a factor of 1,000 with respect to solar or cosmic abundances, and carbon was depleted by a factor of 4 in the gaseous fraction. The results of the 1986 study of Comet Halley confirmed the average chemical model and showed that the carbon missing in the gas was actually present in the dust. Except for hydrogen (and presumably helium), it appears that all elements are roughly in cosmic proportions in comets in spite of their extremely low gravity (10^{-4} times that of the Earth). This emphasizes the pristine nature of comets. Unlike most bodies of the solar system, comets obviously have never been severely processed by any heating episode since their formation. If the accretion of comets occurred at very low temperatures, near absolute zero, the water ice in a newly formed comet must be amorphous. Idealized models show that the transition to cubic ice might be the cause of sudden flare-ups between 3 and 6 AU.

ORIGIN AND EVOLUTION OF COMETS

Evolution of orbits

All observed comets make up an essentially transient system that decays and disappears almost completely in less than one million years. Since they all pass through the solar system, planetary perturbations eject a fraction of them into deep space on hyperbolic orbits and capture another fraction on short-period orbits. In turn, those that have been captured decay rapidly in the solar heat. Fortunately, there is a permanent source of new comets that maintains the steady state—namely, the outer margin of the Oort cloud. As explained above, these so-called new comets are those Oort-cloud comets whose perihelia have been brought down into visibility—*i.e.*, into the inner planetary system where they display their spectacular decay through comas and tails. Comets within the bulk of the Oort cloud are unobservable, not only because they do not develop comas and tails but also because they are too far away.

Formation of the Oort cloud. Any modern theory about cometary origins must first explain the origin of the Oort cloud. None of the comets observed today left the Oort cloud more than three or four million years ago. The Oort cloud is, however, gravitationally bound to the solar system, which it follows in its orbit around the Milky Way Galaxy. Therefore, it is likely that the Oort cloud has existed for a long time. The most probable hypothesis is that it was formed at the same time as the giant planets by the very process that accreted them. The Soviet astronomer Viktor S. Safronov developed this accretionary theory of the planetary system mathematically in 1972. According to his model, the planets originated from a disk or a ring of dust around the Sun, and cometary nuclei are noth-

The Oort cloud as a by-product of the accretion of the giant planets

ing more than primordial planetesimals that accreted first and became the building blocks of the planets. From the accreted mass of the giant planets, Safronov predicted the correct order of magnitude of the mass of the Oort cloud, which was built up by those planetesimals that missed colliding with the planetary embryos and were thrust far away by their perturbations. In effect, the Oort cloud in this theory becomes the necessary consequence and the natural by-product of the accretion of the giant planets.

Later in the 1970s the American astronomer A.G.W. Cameron developed a much more massive model of the protostar nebula, in which the comets accreted in a circular ring at some 1,000 AU from the Sun, which is far beyond the present limits of the planetary system. The primeval circular orbits were then transformed into the elongated ellipses present in the Oort cloud by mass loss of the primitive solar nebula. Both the Cameron and Safronov models put the origin of comets together with that of the solar system some 4.6 billion years ago. Plausibility is given to the general idea of accretion from dust disks by the existence of such disks around many young stars—a fact established by infrared observations in the 1980s and confirmed visually in at least one case (β Pictoris). Further support is found in clues derived from meteorites.

Since the early 1980s, new ideas have been explored to determine whether the Oort cloud could be much younger than the solar system or at least periodically replenished. The role of the massive and dense molecular clouds that exist in interstellar space has been reexamined in different ways. Could comets have accreted in these clouds directly from interstellar grains? Mechanisms for later capturing them into the Oort cloud cannot be very effective, but the efficiency is not capital, and some possibilities have been proposed. Since the solar system itself was probably formed from the gravitational collapse of such a molecular cloud, it seems more likely that either comets or the interstellar grains that were going to accrete into comets followed suit during gaseous collapse and were put into the Oort cloud at the same time that the planets were being formed. Elemental isotopic ratios deduced from the Comet Halley flyby have not brought about any conspicuous anomalies that could be attributed to matter coming from outside the solar system. So far, observational clues all favour the idea of cometary matter deriving from the same primeval reservoir as the stuff of the solar system, but it must be recognized that the evidence remains weak.

Possible pre-solar-system origin of comets. Telltales based on the chemical constitution of cometary nuclei as well as on the evolution of their orbits suggest that the origin of comets goes back beyond that of the planets and their satellites. Two scenarios are among the likeliest possibilities. In the first, comets had already accreted in all dense molecular clouds of the Milky Way Galaxy by the agglomeration of interstellar grains covered by a frost of organic molecules that cemented them together. Later, such a cloud collapsed to form the solar system. In the second scenario, dense molecular clouds were not able to accrete their frosty interstellar grains into larger bodies. When one of these molecular clouds collapsed to form the future solar system, however, the interstellar grains did likewise and eventually formed a dusty disk around the central star—the proto-Sun. Accretion into objects of 10-kilometre diameter is more likely in dusty disks of this type. The outer grains of the disk had not lost their frost, and some of them were ejected into the Oort cloud during the accretion of planetesimals into giant planets after some very moderate processing by heat. It is hoped that one day, space probes will secure data that will make it possible to determine whether frosty interstellar grains have lost their identity or can still be recognized as pristine and unaltered objects in cometary dust.

Comets seem to be the most pristine objects of the solar system, containing intact the material from which it was formed. Included are the hydrogen, carbon, oxygen, nitrogen, and sulfur atoms needed to build the volatile molecules present in the terrestrial biosphere (including the oceans and the atmosphere). Comets also seem to be the link between interstellar molecules and the most primitive meteorites known—the carbonaceous chondrites (see below *Types of meteorites*). The molecules required to initiate prebiotic chemistry (*e.g.*, hydrogen cyanide, methyl cyanide, water, and formaldehyde) are present in interstellar space just as they are in comets; larger prebiotic chemistry molecules (*e.g.*, amino acids, purines, and pyrimidines) occur in some chondrites and possibly in comets. An early cometary bombardment of the Earth, predicted in some accretion models of the solar system, may have brought the oceans and the atmosphere, as well as a veneer of the molecules needed for life to develop on the Earth. Comets could well be the link between interstellar chemistry and life. (A.H.De.)

Meteoroids, meteors, and meteorites

METEOROIDS

General considerations. The solar system contains many small bodies that move in orbits sufficiently eccentric to cross over and intersect the orbit of the Earth. When their orbits do intersect that of the Earth, the probability of the objects colliding with the planet becomes quite high. A body of this kind entering the Earth's atmosphere is called a meteoroid. Such bodies range from small particles less than one micrometre in size (about 10^{-12} gram in mass) up through objects several centimetres or metres in diameter and grade into kilometre-size bodies large enough to be observed as astronomical objects through telescopes. They enter the upper atmosphere with velocities of 11 to 72 kilometres per second. Interaction with the atmosphere heats incoming objects that are larger than about 0.01 centimetre to temperatures high enough to cause them to become incandescent, vaporize, and heat the surrounding air. As a result of this sequence, such objects are observable from the ground as meteors or, in more popular language, "shooting stars" or "falling stars." Strictly speaking, the term meteor refers only to the phenomena associated with the collision of a meteoroid with the Earth's atmosphere. Scientific usage is not all that strict, however, and the body itself is often called a meteor. Unusually luminous meteors are termed fireballs or bolides.

Variations in size

Owing to its fairly low entry velocity, large mass, and physical strength, a meteoroid sometimes survives its passage through the atmosphere, falls to the ground, and may be recovered as a meteorite. Recovered meteorites, ranging in mass from a few grams to several tons, are often exhibited in museums. Most meteorites either consist of rocky—chiefly silicate—material (stony meteorites) or are composed primarily of nickel-iron alloy (iron meteorites; Figures 69 and 70). In stony-iron meteorites, massive nickel-iron alloy is intermixed with silicate material.

In addition to these relatively large meteorites, it is possible to recover much smaller objects about 0.001 centimetre in diameter called micrometeorites on filters attached to aircraft flying in the stratosphere. These micrometeorites (often referred to as interplanetary or cosmic dust) also accumulate on the bottom of the deep ocean. The larger ones can be identified and separated from cores drilled from the muddy deposits on the seafloor.

Micrometeorites

The largest meteoroidal bodies are observable through telescopes as astronomical objects. These include the Apollo objects, bodies of asteroidal appearance with diameters ranging from a few hundred metres to several kilometres that come closer to the Sun than to the Earth's orbit. There are about 700 of these larger than one kilometre in diameter. Because comets can strike the Earth, they, too, can be thought of as large meteoroidal bodies.

When meteoroids are sufficiently large (*i.e.*, 100 metres to several kilometres in diameter), they can pass through the atmosphere without slowing down appreciably. When they strike the Earth's surface at velocities of many kilometres per second, the kinetic energy released is sufficient to produce an impact crater. In many ways such craters resemble those produced by nuclear explosions. They are often called meteorite craters, in spite of the fact that the impacting meteoroids themselves are almost entirely vaporized during the explosion. High-velocity impact by objects of this kind on the Moon, Mercury, and Mars are in large part responsible for the heavily cratered appearance of the surface of these bodies. The cratering record

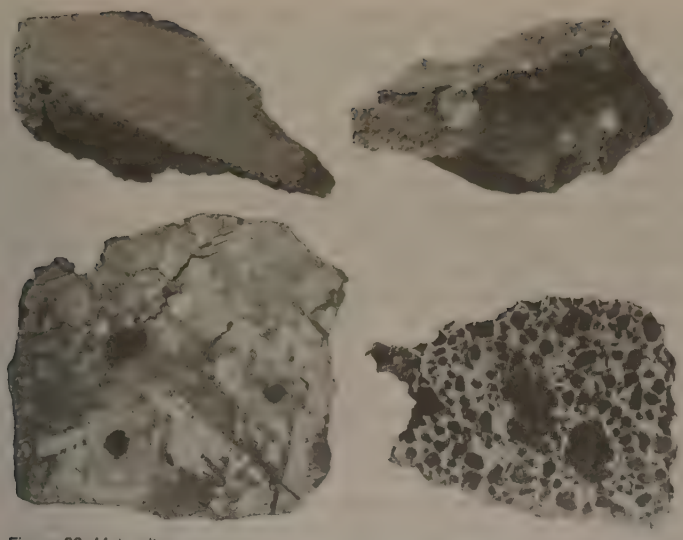


Figure 69: *Meteorites.*
(Top left) Ankober (Ethiopia) ordinary (H4 olivine-bronzite) chondrite (reduced). One surface has been sawed and polished, revealing the internal structure. The white spots are reflections from nickel-iron metal; the surrounding gray material is composed of silicate minerals. (Top right) Allende carbonaceous (CV3) chondrite (reduced 3.8×). The large white inclusions consist primarily of refractory aluminum- and calcium-bearing minerals, which are embedded in a dark gray matrix containing fine-grained minerals formed at much lower temperatures. (Bottom left) Sawed, polished, and acid-etched interior surface of the Osseo (Ontario) iron meteorite (coarsest octahedrite; reduced 4.2×). This treatment reveals the "Widmanstätten pattern" resulting from coarsely crystalline kamacite (α-iron-nickel alloy). (Bottom right) Polished and etched section of the Salta stony-iron (pallasite), composed of roughly equal amounts of olivine (magnesium-iron silicate in the form of dark grains) and nickel-iron alloy (the shiny crystals; reduced 4.1×).
(Top left) J.A. Wood, (others) Smithsonian Institution

on the Earth and Moon also shows that there are many meteoroids in the intermediate mass range between the larger recovered meteorites (a few metres in diameter) and the Apollo objects.

Relationship of meteoroids to asteroids and comets. Most of the mass of the solar system resides in its larger bodies, the Sun and the planets. The planets move about the Sun in stable and well-separated orbits. It is almost certain that these orbits and thus the positions of the planets have undergone only minor changes since the formation of the solar system some 4.6 billion years ago. In addition, the planets are large enough to retain on their surfaces nearly all the debris produced by impact craters.

On the other hand, a smaller fraction of the mass of the solar system is found in bodies of such small size or in orbits so eccentric that their physical survival or orbital stability has been in jeopardy throughout the history of the solar system. Most of these bodies are found in either of two regions of the solar system: the asteroid belt, between the orbits of Mars and Jupiter (mostly between 2.2 and 3.4 AU), and the cometary, or Oort, cloud extending far beyond the orbits of Neptune and Pluto to distances of more than 10,000 AU.

G. Kurat

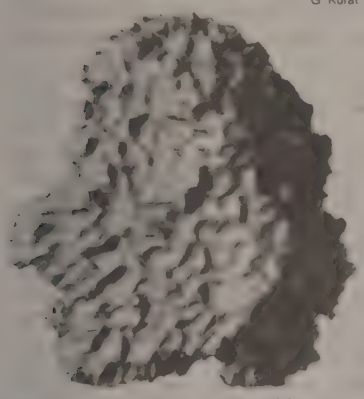


Figure 70: Cabin Creek (Arkansas) iron meteorite (medium octahedrite; reduced 8.1×). The dimpled appearance of the face of the meteorite was caused by melting of its surface upon entering the Earth's atmosphere.

The members of the asteroid belt have high enough eccentricities and inclinations that they collide with one another at velocities averaging about five kilometres per second. Because of this, it is unlikely that most asteroids larger than about 75 kilometres in diameter have survived collisional destruction over the entire 4.6 billion-year history of the solar system. The present-day smaller asteroids are debris formed by fragmentation of larger asteroids caused by this natural grinding process in the asteroid belt. The grinding extends down to bodies the size of meteorites exhibited in museums and even to much smaller particles in the form of fine dust.

Fragmentation of asteroids

Some of this collisional debris is orbitally unstable. The same collisions that produced the fragments propel a portion of them into chaotic orbits controlled by dynamical resonances with the motions of the planets. As a result of collisions and orbital instability, asteroidal material is injected into orbits that cross the Earth's orbit. Some of this matter collides with the atmosphere of the Earth and becomes visible meteors. A small number of the particles survive as meteorites.

The much more distant cometary bodies also are vulnerable to destruction. Their orbits extend to the threshold of interstellar space. Passing stars and interstellar molecular clouds gravitationally perturb some of these bodies into orbits with perihelia within several astronomical units of the Sun. A small fraction of the comets that are perturbed into orbits approaching the Sun experience close approaches to planets, particularly Jupiter, and this leads to further orbital perturbations. In some cases, this causes the orbital eccentricity and orbital period of a comet to be drastically reduced to periods from about three to several hundred years.

Comets are small bodies consisting of a mixture of ice and less volatile material, the latter mostly in the form of dust (see above *The nature of comets*). When a comet is perturbed into an orbit that comes within a few astronomical units of the Sun, the ice begins to evaporate, giving rise to the luminous object commonly thought of as a comet. Dust is swept away from the comet along

with the evaporating ice. Most of the fine dust escapes the Sun's gravitational field and is lost to interstellar space. Some of the dust, however, is in larger particles typically millimetres to centimetres in size. As compared with the finer particles of dust, these larger bodies have a smaller ratio of surface area to mass, and this renders them less susceptible to the solar radiation forces, proportional to the cross-sectional area, that drive the finest dust out of the solar system. Thus, these larger cometary particles can remain in short-period orbits and can intersect the orbit of the Earth. When they collide with the atmosphere, meteoroids of this kind also appear as meteors.

Association of meteor showers with meteoroids of cometary origin

The best evidence for the cometary origin of certain meteoroids is identification of the orbits of meteoroids with those of known periodic comets (see above *Modern cometary research*). It is possible to determine these orbits by photographing the meteors with special cameras as they pass through the Earth's atmosphere. This identification has been made for several of the meteors that produce meteor showers. In this way inferences can be made about the physical properties of particulate cometary material.

The asteroid belt and the outermost part of the solar system should be thought of as long-lived though somewhat leaky reservoirs of meteoroidal bodies. They are long-lived enough to retain a significant quantity of primordial solar system material but leaky enough to permit the escape of the observed quantity of Earth-crossing material. This quantity of Earth-crossing material represents an approximate steady-state balance between the input from the storage regions and the loss by ejection from the solar system, collision with the Earth, the Moon, and other planets, or vaporization by impact.

Evidence for meteoroids of asteroidal origin: radiometric ages of meteorites. The most meaningful way of describing and classifying meteoroids would be in terms of their various sources in the solar system. To do so, it is necessary to find criteria that discriminate among these sources.

Unlike the association of some meteoroids with the actual observed orbits of comets, direct orbital identification is not possible for meteoroids of asteroidal origin. The asteroids in the asteroid belt that are the ultimate parents of these meteoroids are not in Earth-crossing orbits. A meteoroid entering the atmosphere must be in an Earth-crossing orbit. Therefore, the observed meteoroid orbit cannot be the same as that of its source. The evolution of the asteroidal collision debris from the asteroid belt is complex, involving collisional fragmentation and major orbital changes caused by planetary gravitational perturbations. Although this evolution is quantitatively understandable in terms of known physical processes, the complex history destroys the information required to directly link the meteoroid with its source. Unlike cometary meteors, there is, for example, no good evidence that asteroidal meteors occur in streams, nor would such clustering of orbits be expected on theoretical grounds. The link between asteroidal meteoroids and their sources must be established in ways that are less direct.

This link is supplied by laboratory investigations of meteoroids that have survived passage through the atmosphere—namely, meteorites. As a result of these investigations, it is generally believed that nearly all recovered meteorites are fragments of asteroids. A small fraction, about 0.5 percent, may be of lunar or Martian origin. It is possible, but not at all firmly established, that some meteorites may be derived from comets.

The reasoning that leads to identification of most meteorites as asteroidal meteoroids is fairly complex but quite compelling. Perhaps the best starting point would be that based on studies of the daughter isotopes produced by the radioactive decay of radioactive parent isotopes of such elements as uranium, thorium, rubidium, potassium, and samarium. As is commonly done for terrestrial rocks, the ages of meteorites can be determined by isotopic measurements of this kind.

Isotopic measurements

When this is done, all but a very few meteorites show good evidence of having been formed, from the mineralogical point of view, some 4.6 billion years ago. Similar techniques have been used to infer, somewhat less directly, the age of the Earth and the Moon, with the same result. Within an uncertainty of less than about 100 million years, this age represents the time in the past when the Sun, the planets, and their natural satellites, along with the asteroids and comets, formed from interstellar dust and gas. Because of geologic processes that produced younger rocks on the Earth and Moon, their evidence for this primordial age of formation has been obscured but not erased. By contrast, the evidence for the age of the solar system is clearly preserved in meteorites.

Although present-day meteorites are the products of a long series of events that caused physical fragmentation, in most cases the mineral grains of which they are composed have undergone little chemical alteration since the early years of the solar system. As a consequence, the isotopic ratios from which the ages are calculated preserve the evidence for this very ancient time of formation.

If meteorites are asteroidal fragments, this is what would be expected. Theoretical calculations of the thermal evolution of bodies as small as asteroids predict that they would not exhibit the long-continuing history of igneous and metamorphic rock formation characteristic of large bodies such as the Earth and Moon. As in the case of the latter, however, the decay of radioactive isotopes produces heat in the interiors of asteroids. Yet, because of their small size, this heat is readily conducted to the surface and radiated to space. The internal temperature thus would never rise high enough to produce the metamorphism or melting that would chemically alter the ancient mineralogy. In the much larger Earth, on the other hand, the heat is not efficiently transported to the surface by convection until the temperature increases to near its melting point.

By itself, this preservation of ancient mineralogy does not, however, clearly indicate that the meteorites are of asteroidal rather than of cometary origin. Comets are even smaller bodies than asteroids and would be heated even less. If, in addition to the volatile material and fine dust emitted by comets as they approach the Sun, there exists ancient rocky material in their nuclei, one would expect the primordial isotopic age record to be preserved in cometary meteoroids as well.

Distinguishing between asteroidal and cometary sources

The clue to the distinction between asteroidal and cometary sources for at least the most common meteorites is provided by minor mineralogical disturbances that slightly alter the otherwise clear preservation of the primordial isotopic age record in meteorites. Such disturbances are attributable to metamorphism associated with evidence for collisional shocks preserved in the detailed mineralogy of meteorites. For example, many representatives of one very common type of chondritic meteorite, the hypersthene chondrite, display glass veins and other evidence of shock-induced metamorphism of the sort that would be expected to result from collisions between asteroids. (Chondrites, the most common type of stony meteorite, are so called because they contain small primordial silicate spherules known as chondrules.) Ages measured by the decay of potassium to argon show that this shock metamorphism took place quite late in the history of the solar system—about 500 million years ago. Ages measured by other isotopic decay systems on these same meteorites still display the more common value of 4.6 billion years (Figure 71). Furthermore, these shocked meteorites exhibit a very strong chemical and textural kinship to unshocked meteorites that record completely the isotopic age of the solar system. It is not plausible that such similar rocks should be provided by sources from regions as far removed from one another as the asteroid belt and the cometary cloud. Other evidence for impact events late in solar system history is provided by dating shocked fragments of meteoritic material embedded in meteorites of different types which are presumably a result of collisions in space. Quantitative comparison of the collisional history expected for the sparsely populated, low-relative-velocity source regions of comets with that expected for the more densely populated, higher-velocity asteroid belt argues strongly for the asteroid belt being the actual site of these collisions and thus the source region for the meteorites.

Such conclusions are supplemented by observational and theoretical dynamical studies of meteorite orbits. The results of these studies are in agreement with an asteroidal

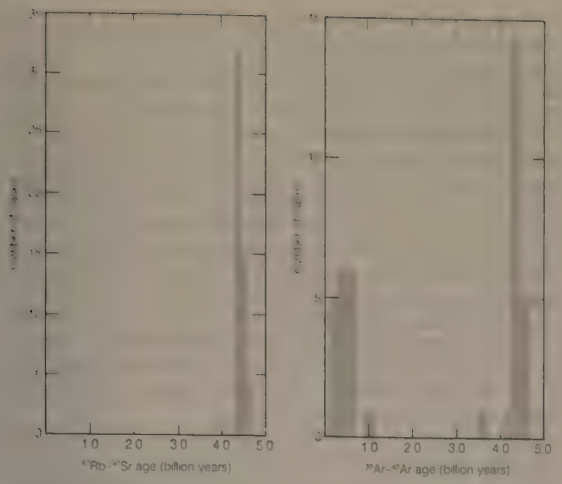


Figure 71: *Distribution of formation and metamorphism ages of ordinary chondrite meteorites.*
(Left) Ages measured by the decay of ^{87}Rb to ^{87}Sr. The "isotopic clocks" of these meteorites were set by heating events essentially at the beginning of solar system history. (Right) Ages based on the decay of ^{40}K to ^{40}Ar. Many of these ages also cluster around 4.6 billion years, but a significant number have been "reset" by collisionally induced shock metamorphism on the asteroid from which the meteorite was fragmented.

(Left) Based on data from J.F. Minster (1979) and a series of papers by K. Gopalan, S.K. Kaushal, and G.W. Wetherill in *Journal of Geophysical Research* (1968–71), copyright by the American Geophysical Union. (Right) Data compiled by J.T. Wasson (1985) based on data from D. Bogard, *Journal of Geophysical Research*, vol. 81, p. 5664, (1976), copyright by the American Geophysical Union; G. Turner, *Meteorite Research* (1969), and G. Turner *et al.*, *Proceedings of the Ninth Lunar Planetary Scientific Conference* (1979)

origin for at least the most abundant types of meteorites but are not in agreement with a cometary origin.

Meteoroids of less certain origin. Identification of many meteoroids with either asteroids or comets can be securely established. It is then possible to extend these identifications by extrapolation to cases where the distinction is less obvious. There are, for example, many shower meteors in meteor streams for which no comet is presently identifiable; however, the orbits and characteristics of their passage through the atmosphere are very similar to those of stream meteors that are associated with known comets. Even when such meteors are not in streams, their orbits with aphelia and physical characteristics similar to those of cometary stream meteors leave little room to question their cometary parentage. It also has been possible to identify a number of very bright meteors (fireballs) with asteroidal meteorites by comparing the details of their atmospheric flight with those of recovered meteorites (see below *Fireball networks*).

There are many meteoroids whose asteroidal or cometary nature remains highly uncertain. These include objects that are presently in orbits very similar to identifiably asteroidal meteoroids but that exhibit in their atmospheric flight the greater fragility associated with meteoroids known to be derived from comets. Are these fragments of comets that have been perturbed into orbits traditionally associated with asteroids, or do they represent a type of asteroidal material that is too weak to survive as a meteorite? Answers to such questions are needed if scientists are to use the information preserved in primitive bodies—comets and asteroids—in the most effective way. There is much more that can be learned from further observational studies of meteors, laboratory investigations of meteorites, Earth-based observations of comets and asteroids, and spacecraft missions to these small primitive bodies. Continuation of this work holds much promise for clarifying the processes by which both the Earth and the solar system came into existence.

In addition to cometary and asteroidal sources, a very minor but identifiable contribution to the meteoroidal population is derived from the Moon. Such particles are ejected from primary cratering of the lunar surface by large meteorites. By comparison with lunar rocks returned by the U.S. Apollo missions, several small meteorites collected on the Antarctic ice sheet have been positively identified as lunar meteorites. It also appears probable that certain relatively rare types of meteorites (shergottites, nakhlites, and chassignites) are fragments of igneous rocks from Mars. Some day evidence for meteoroids from Mercury, Venus, the Earth, and the interstellar medium may be identified.

METEORS

Characteristic observational effects. On any clear night in the countryside beyond the bright lights of cities, one can observe with the naked eye several meteors (or shooting stars) per hour as they streak through the sky, with durations ranging from a small fraction of a second up to several seconds. Quite often they vary in brightness along the path of their flight, appear to emit "sparks" or flares, and sometimes leave a luminous train that lingers after their flight has ended. These meteors are the result of the high-velocity collision of meteoroids with the Earth's atmosphere. Nearly all such interplanetary bodies are small fragments derived from comets or asteroids.

The observed apparent brightness of these easily observable meteors covers the same range of brightness as the stars visible to the unaided eye (*i.e.*, from about zero to fifth astronomical magnitude). They constitute a portion of an Earth-impacting interplanetary flux of similar bodies ranging in mass from less than one nanogram up to millions of tons. The smaller bodies are too faint to be seen with the naked eye but are observable with the aid of binoculars and telescopes or by radar reflection. Brighter meteors—of magnitudes ranging in brightness from that of Venus (−4 magnitude) to greater than that of the full Moon—are less common but are not really unusual. These are produced by meteoroids with masses ranging from several grams up to about one ton.

The brightest meteor (possibly of cometary origin) for which historical documentation exists struck on June 30, 1908, in the Tunguska region of central Siberia and rivaled the Sun in brightness. The energy delivered to the atmosphere by this impact was roughly equivalent to that of a 10-megaton thermonuclear explosion and caused the destruction of forest over an area of about 2,000 square kilometres. The geologic record of cratering attests to the impact of much more massive meteoroids, including objects with kinetic energies equivalent to 100 million megatons. Fortunately, impacts of this magnitude occur only once or twice every 100 million years. It is hypothesized that large impacts of this kind may have played a major role in determining the course of biological evolution by causing simultaneous mass extinctions of many species of organisms, possibly including the dinosaurs some 65 million years ago. If so, the replacement of reptiles by mammals as the dominant land animals, the eventual consequence of which was the rise of the human species, would be the result of a grand example of a phenomenon observable every clear night.

Heights and velocities

The visibility of meteors is a consequence of the high velocity of meteoroids in interplanetary space. Before entering the region of the Earth's gravitational influence, their velocities range from a few kilometres per second up to as high as 72 kilometres per second. As they approach the Earth, within a few Earth radii, they are accelerated to even higher velocities by the planet's gravitational field. As a consequence, the minimum velocity with which a meteoroid can enter the atmosphere is equal to the Earth's escape velocity of 11 kilometres per second. Even at this minimum velocity, the kinetic energy of a meteoroid would be 6×10^4 joules per gram of its mass. This can be compared with the energy of about 4×10^3 joules per gram produced by chemical explosives, such as TNT. As the meteoroid is slowed down by friction with atmospheric gas molecules, this kinetic energy is converted into heat. Even at the low atmospheric density at altitudes of 100 kilometres (6×10^{-10} gram per cubic centimetre compared with 10^{-3} gram per cubic centimetre at sea level), this heat is sufficient to vaporize and ionize the surface material of the meteoroid and dissociate and ionize the surrounding atmospheric gas as well. Electronic transitions effected by this excitation of atmospheric and meteoroidal atoms produce a luminous region, which travels with the meteoroid

and greatly exceeds its dimensions. At deeper levels in the atmosphere, a shock wave may be produced in the air ahead of the meteoroid. This shock wave interacts with the solid meteoroid and its vapour in a complex way. About 0.1 to 1 percent of the original kinetic energy of the meteoroid is transformed into visible light.

Ablation and fragmentation

This great release of energy destroys meteoroids of small mass—particularly those with relatively high velocities—very quickly. This destruction is the result both of ablation (the loss of mass from the surface of the meteoroid by vaporization or as molten droplets) and of fragmentation caused by aerodynamic pressure that exceeds the crushing strength of the meteoroid. For these reasons, numerous meteors end their observed flight at altitudes above 80 kilometres, and penetration to as low as 50 kilometres is unusual. Nevertheless, some meteoroids survive to much lower altitudes owing to a combination of relatively low entry velocity (< 25 kilometres per second), large mass (>100 grams), and fairly high crushing strength ($>10^7$ dynes per square centimetre). Those that are recoverable as meteorites lose their kinetic energy before the meteoroid is completely destroyed. They are effectively stopped by the atmosphere at altitudes of 5 to 25 kilometres. Following this atmospheric braking, they begin to cool, their luminosity fades, and they fall to the Earth at low terminal velocities of 100 to 200 metres per second. This "dark flight" of the meteoroid may be several minutes in duration, in contrast to the few seconds of visible flight.

The passage of meteoroids through the atmosphere produces atmospheric shock waves that penetrate to the ground. The penetration of a meteoroid in the kilogram range to altitudes of about 40 kilometres can thereby produce sounds on the ground similar to sonic booms or thunder. These sound waves can be intense enough to become coupled to the ground and recorded by seismometers.

The effect of the final impact with the ground of meteorites in the kilogram mass range could be considered an anticlimax. The fall can go unnoticed by those near the impact site, the impact being signaled only by a whistling sound and a thud. For this reason, many meteorites are recovered only because at least one fragment of the meteoroid strikes a house, drawing the attention of the residents to an unusual event.

Orbits of meteoroids and meteor showers. Prior to entering the gravitational field of the Earth, a meteoroidal body, like all bodies of the solar system, moves around the Sun in an elliptic Keplerian orbit. If this orbit can be determined, valuable information relevant to identifying the source—the parent body—of the meteoroid can be obtained.

When the coordinates in the sky of the trajectory of a meteor are observed from two or more well-separated stations, the direction in which the meteoroid was moving in space before it encountered the Earth can be estimated reasonably well by triangulation. This direction is called the radiant of the meteor. If the motion of the meteoroid is thought of as a velocity vector, such observations determine approximately the direction of this vector. To determine the meteoroid's orbit, however, requires ascertaining not only the direction but also the magnitude of the velocity vector. Although well-trained visual observers are able to estimate the coordinates of the meteor trajectories and thereby determine radiants fairly well, their velocity estimates have proved to be too uncertain to be useful for orbit determination.

This problem was overcome during the 1940s by the introduction of astronomical cameras specially designed for studying meteors. These wide-field cameras were equipped with a rotating shutter that periodically interrupted the light to the photographic plate. The shutter breaks permitted calculation of the speed of a meteor along its path. The position of the meteor's trajectory with respect to the stars photographed on the same plate also was measured accurately. Such observations made at two or more stations could then be used to calculate precisely the orbit of the meteoroid before it encountered the Earth. During the 1940s, special radar instruments also were applied to the study of meteors generally fainter than those observed photographically.

"Showers" of meteors have been known since ancient times. On rare occasions, these showers are very dramatic, with thousands of meteors falling per hour. More often, the background hourly rate of roughly 5 observed meteors increases up to about 10–50. Shower meteors characteristically have nearly the same radiant. This means they are all moving in the same direction in space. As a consequence, plots of meteoroid trajectories on a star map converge at a single point, the radiant, for the same reason that parallel railroad tracks appear to converge at a distance. The new photographic data fully confirmed the belief that meteors belonging to a particular shower had not only the same radiant but similar orbits as well. In other words, the meteoroids producing the meteor showers move in confined streams around the Sun. The introduction of radar observation led to the discovery of several new meteor streams that were totally invisible to cameras because they came from radiants in the daytime sky.

Radiants of shower meteors

A fact of great importance, fully confirmed by the photographic data, is the association of many meteor streams with the orbits of observed comets. A list of the more important meteor-stream orbits and associated comets is given in Table 26. The streams with cometary associations represent debris ejected from a comet along its orbit through space. A recently formed stream, the Leonids, tends to appear in great strength every 33 or 34 years, the same as the period of the parent comet, Temple-Tuttle. These meteoroids are clustered in a compact swarm moving in the orbit of the comet. With the passage of 1,000 years or so, the slightly different orbits of the meteoroids will cause them to disperse more uniformly along the orbit of the comet. In such cases, a shower, usually weaker, occurs annually when the Earth's orbit intersects the orbital plane of the meteor stream. After a still longer period, about 10,000 years, planetary perturbations will cause the orbits of the stream meteoroids to disperse into different orbits, and their identity as members of a stream will gradually disappear.

Meteors that do not appear to belong to streams are called sporadic. It is likely that in some sense all meteoroids are, or have been, stream members because the physical processes that release meteoroids from either comets or asteroids do so in great numbers. Sporadic meteors are therefore the result of streams too weak to be distinguished from one another or old streams so dispersed as to be no longer recognizable.

Sporadic meteors

There is one rather strange example of a major meteor shower that is clearly identifiable with an astronomical object that, at least at first glance, does not appear to be a comet: the Geminid shower and 3200 Phaeton, respectively. The latter (formerly designated 1983TB) exhibits none of the usual cometary features, a nebulous halo and tail; it simply looks like a small Earth-crossing asteroid. If the stream is cometary, it means that a comet that produced meteoroids prolifically only a few thousand years ago has now completely ceased its cometary activity and looks more like an asteroid. The orbits of 3200 Phaeton and the Geminids also are unlike those of comets in that their aphelia are at 2.4 AU, well within the orbit of Jupiter.

Table 26: Principal Visually Observable Meteor Showers

shower	average date of maximum	normal duration (days)	visual strength (Northern Hemisphere)	entry velocity (km/sec)	associated comet
Quadrantid	January 3	1	medium	41	not known
Lyrid	April 22	1	irregular	48	1861 I (Thatcher)
Eta Aquarid	May 3	5	weak	66	Halley
S. Delta Aquarid	July 29	8	medium	41	not known
Capricornid	July 30	3	medium	23	not known
Perseid	August 12	5	strong	59	Swift-Tuttle
Andromedid	October 3	11	weak	21	Biela
Draconid	October 9	1	irregular	20	Giacobini-Zinner
Orionid	October 21	2	medium	66	Halley
Taurid	November 8	30	weak	28	Encke
Leonid	November 17	<1	irregular	71	Temple-Tuttle
Geminid	December 14	4	strong	34	3200 Phaeton*

*This body exhibits no cometary activity and is possibly of asteroidal rather than of cometary origin.
Source: Data derived primarily from A.F. Cook in *NASA SP-319* (1973).

If 3200 Phaeton came from the Oort cloud of comets (see above *Comets*), it must at one time have crossed the orbit of Jupiter. Over a span of millions of years, it is not out of the question that close encounters with the Earth and Venus could gravitationally perturb a Jupiter-crossing orbit into an orbit of this kind. The observed rate at which matter is lost from comets, however, seems to indicate that their inventory is exhausted in only a few thousand years. Thus, the millions of years required for this orbital evolution does not appear to have been available.

It could be hypothesized that 3200 Phaeton was never a comet at all but simply an Apollo object that strayed from the asteroid belt by the reasonably well-understood resonant perturbations that can cause this to occur. The Geminid meteors would then be explained as fragments produced by an asteroidal collision while 3200 Phaeton is traversing the portion of its orbit that is in the asteroid belt. There are serious problems with this explanation. Quantitative calculations show that an asteroidal collision of the required magnitude during an interval of only a few thousand years is very unlikely. Studies of the historical orbital evolution of the Geminid stream suggest that its orbit is incompatible with a single outburst of meteoroids; it is more like one expected from a body that produced a series of outbursts. Finally, the orbit is not the kind one would expect for an Apollo object perturbed from the asteroid belt. Its perihelion is too close to the Sun. An understanding of the mystery of 3200 Phaeton and the Geminids is likely to contribute much to the understanding of comets, the origin of meteor streams, and the relationship between Apollo objects, comets, and asteroids.

Fireball networks. A very significant development in meteor science occurred during the 1960s. This was the establishment of large-scale networks for photographing very bright meteors, or fireballs. These networks were designed to provide all-sky coverage of meteors over areas of about a million square kilometres. Three such networks were developed: the Prairie Network in the central United States, the MORP (Meteorite Observation and Recovery Project) network in the prairie provinces of Canada, and the European Network with stations in Germany and Czechoslovakia. The most complete set of published data is that of the Prairie Network, which was operated by the Smithsonian Astrophysical Observatory from 1964 to 1974.

Principal objectives of the networks

An original goal of these networks was to recover a larger number of meteorites for laboratory studies. Other objectives were to determine the orbits of the recovered meteorites and to compare the inferences of meteor theory regarding the density and strength of meteoroids with "ground truth" provided by the study of the same meteoroids in the laboratory. The goal of recovering meteorites had only limited success. Three meteorites were recovered, one by each of the networks. All three meteorites were ordinary chondrites, the most abundant type of stony meteorite. During the operation of the networks, many more meteorites were actually recovered by chance collisions with the roofs of houses.

In spite of this meagre record for meteorite recovery, the networks compiled data that became the basis for a new outlook on meteor science and meteoroid sources. Prior to this effort, there was a tendency to regard the study of meteors and meteorites as independent scientific fields that had little to contribute to each other. Meteors were studied by astronomers and were thought to be associated almost entirely with fragile and low-density "dust balls." On the other hand, meteorites were dense rocks studied by geochemists in the laboratory as samples of the primordial solar system. Little thought was given to why meteor astronomers did not concern themselves with meteorites.

Straightforward application of conventional meteor physics to determine the density of the three recovered meteorites led to the incorrect conclusion that these dense rocks were also low-density objects. This clearly showed that there was something wrong with meteor physics as traditionally applied. A likely, but still not proven, explanation was that the value of luminous efficiency, conventionally used to relate the mass of a meteoroid to the brightness of a meteor, was too low. As a result, the mass of the meteoroid calculated from the luminosity of the meteor was too high. When this large "photometric" mass was combined to measure the cross-sectional area of the meteoroid (using the rate at which it was observed to slow down by atmospheric gas drag) and thereby its radius and volume, a spuriously low density was obtained.

Impact of photographic data on meteor studies

The photographic data from the three fireballs recovered by the networks permitted a more direct empirical approach to the analysis of meteor data. It was found that the atmospheric trajectories of the recovered meteorites, including the end height at which they ceased to be luminous, could be accurately reproduced if the "dynamic mass," determined by the deceleration of the meteor, were used in the theory instead of the photometric mass. It also was found that the ratio of the photometric mass to the dynamic mass was a constant. Laboratory measurements of cosmic-ray effects on the recovered meteorites led to a calculation of the "true mass," which was intermediate between the photometric mass and the dynamic mass. Finally, the light curve (the plot of brightness versus altitude) was similar for the three meteorites.

These results, obtained from the recovered meteorites, could then be used to identify similar objects in the other fireball photographs. Their presence certainly could be expected, because meteorites are produced by fragmentation processes in space similar to those studied in the laboratory. Both experimental and theoretical studies of these processes demonstrate that for every large fragment there must be many small ones.

The recovered fireballs were among the very brightest observed by the photographic networks. Accordingly, there were among the fireball data many objects physically identical to the recovered meteorites. In short, the problem of determining which fireballs were meteorites no longer was dependent on uncertain first principle measurements of density. The empirical data obtained from the recovered meteorites could be used to check the record of each individual fireball, testing quantitatively whether or not the object "looked like a meteorite."

To date, about 30 fireballs have been identified as stony meteorites in this way. The adoption of this approach has increased scientific knowledge of the distribution of meteorite orbits by an order of magnitude.

Application of the same method of analysis shows that fireballs from the Taurid shower, associated with Comet Encke, do not look like meteorites, or at least not like ordinary chondrites. On the other hand, they do not resemble dust balls either but appear to have significant physical strength. The stronger objects of this group have a strength comparable to that of ordinary dirt clods. These physical properties overlap with those of some carbonaceous meteorites. Further analysis of existing data can be expected to shed new light on important questions regarding the relationships between meteoroids of various kinds and their sources. If some way could be found to increase the rate at which fireball networks recover meteorites by about an order of magnitude, the empirical approach, proved valuable for identifying ordinary chondrite sources, could be extended to include less abundant types of meteorites.

METEORITES

As noted above, meteorites are meteoroids that survive passage through the Earth's atmosphere. Any source that can eject such material into interplanetary space should therefore, at least in principle, be thought of as a candidate source of meteorites. There is no fundamental reason why all meteorites must come from similar sources.

It turns out, however, that in practice there are some regions in the solar system that are much more effective in introducing material of substantial strength into Earth-crossing orbits than others. Recent laboratory and theoretical studies fully confirm the older belief that most meteorites are fragments of asteroids. These same studies show that a small fraction, less than 1 percent of the meteorites, come from nonasteroidal sources. The lunar origin of several meteorites is well-established, and it is probable that at least eight others come from Mars. There is evidence from fireball data that a small part of the material in cometary orbits (*i.e.*, with aphelia beyond Jupiter)

Asteroidal origin of most meteorites

Table 27: Classification of Undifferentiated Meteorites (Chondrites)

class	group	percentage of observed chondrite falls	total iron (weight %)	Fe—metal / Fe—total (%)	FeO / FeO + Mg (mole %)	chondrules (%)	carbon (weight %)
Ordinary	H	38.1	25–31	58–65	17	80	0.1
	L	46.3	21–23	30–39	22	80	0.1
	LL	8.5	20–23	6–25	27	80	0.1
Enstatite	E	1.8	22–35	70–88	0.05	20	0.4
Carbonaceous	CI	0.7	18–19	0	45	<1	3.1
	CM	2.3	21–24	0.1–0.6	43	2	2.5
	CO	1.0	24–26	3–19	35	70	0.5
	CV	1.3	22–25	0.8–25	35	30	0.5

Sources: Data primarily from B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); and J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

may possess sufficient strength to successfully penetrate the atmosphere. It is not known if any of this material is present in existing meteorite collections. If it is, the best candidate material would be carbonaceous stony meteorites, probably those of type CI (see below), of which five separate falls have been recovered.

With these few exceptions, it is safe to regard all meteorites as samples broken from outcrops of rock or metal, which until fairly recently in solar-system history were part of asteroidal bodies, mostly in the inner region of the asteroid belt (between about 2.2 and 2.6 AU). Like rocks from the Moon, the Earth, or any other similar planetary body, their present state is determined by the total effect of events that occurred on the body throughout the entire history of the solar system. There is no a priori reason why such samples must be pristine samples of a primordial solar nebula from which the present solar system evolved. On the other hand, the principal driving force behind asteroid studies has been the plausible belief that small "primitive" bodies such as asteroids and comets are those most likely to preserve evidence of events that took place in the early solar system. Insofar as this belief is correct, meteorites, samples of these bodies, share this property. Evidence derived from the study of meteorites themselves supports this conclusion.

Types of meteorites. The most fundamental distinction between the various meteorites—no two of which are exactly alike—is the division between chemically undifferentiated and differentiated meteorites. This concept arises from the fact that there is an average chemical composition of the solar system. This average composition must be very close to the composition of the Sun, because the Sun contains most of the mass of the solar system. Spectroscopic comparison of the Sun's chemical makeup with those of other stars shows that its composition in turn is closely related to a cosmic average of the relative abundances of the elements. Important deviations from such average abundances are observed, but these do not invalidate the view that they are deviations from normal abundance ratios determined by the processes by which the chemical elements are formed in stars at various stages of their evolution, returned to the interstellar medium where they are mixed, and then incorporated into new stars and their planetary systems when they are formed.

Since the late 1940s, important advances have been made in the chemical analysis of both meteorites and the Sun. A remarkable result has emerged from this work. Although at one time there appeared to be major differences between the Sun and typical meteorites in the ratios of elements to one another (*e.g.*, iron to silicon), these differences tended to disappear as the accuracy of the measurements improved. It turned out that, for most meteorites and most elements, the solar and meteoritic values of the element ratios relative to silicon (taken as a standard) agreed to within better than a factor of two.

Basic agreement between the meteoritic and solar abundances of the elements

Two kinds of exceptions to this rule were found. Relative to the Sun, the meteorites were deficient in the more volatile elements. For the most volatile elements, hydrogen and the noble gases, the deficiencies were gross—more than a factor of 10,000. For less volatile elements, the deficiencies were smaller; and for the nonvolatile elements, or "refractory" elements, such as iron, magnesium, aluminum, and calcium, the meteoritic and solar abundance ratios were identical within the accuracy of the data.

The other kind of exception to the rule relates to differences between meteorites. For some meteorites, as, for example, those consisting primarily of metallic iron, the similarity between meteoritic and solar abundance ratios fails completely. This also is true for basaltic meteorites, those that appear to have been at one time volcanic magmas and have undergone chemical fractionation of the sort observed in terrestrial igneous rocks.

Undifferentiated meteorites. The meteorites that do obey the rule prove to be of a kind that had already been grouped together on textural grounds—namely, the chondrites. From observed fall rates, this is the most abundant type of meteorite (Table 27). The designation chondrite is based on the occurrence in these meteorites of small (about one millimetre in diameter) spherules called chondrules (Figure 72). In many chondrites, the composition of the chondrules is quite heterogeneous, and the space between

(Top) F. Wlotzka, Max-Planck-Institut fur Chemie, Mainz, Ger., (bottom) J.A. Wood

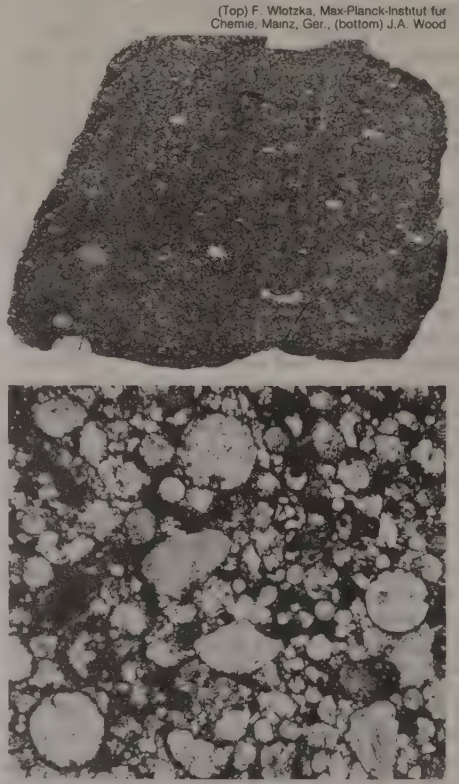


Figure 72: (Top) Sawed and polished section of the Leoville carbonaceous chondrite (CV3; reduced 2.3×). The small, rounded gray objects are chondrules. The larger, whitish objects are refractory inclusions, similar to those seen in Allende (Figure 67, top right). (Bottom) Microscopic view of a thin section of the Tieschitz (Czechoslovakia) ordinary chondrite (olivine-bronzite H3). The round objects are chondrules, some of which have been fractured by collisions after their formation.

the chondrules is filled with a fine-grained matrix material that is richer in volatile elements. In terms of terrestrial rocks, these meteorites seem more akin to sedimentary conglomerates composed of a mechanical mixture of a jumble of components rather than to rocks formed by a process of igneous differentiation.

Not all chondrites contain chondrules of heterogeneous composition. More often, the mixture of heterogeneous chondrules and matrix appears to have undergone a thermal metamorphism that caused the chemical components of the chondrules to come into equilibrium with one another and with the matrix material. This metamorphism was accompanied by the loss of relatively volatile elements. All of these chondrites share one common property. They do not appear to have ever experienced chemical differentiation associated with igneous melting. This shared property is the basis for grouping them all as undifferentiated meteorites even though they clearly differ from one another in the degree to which they have retained volatiles and in various other ways.

Undifferentiated meteorites are classified in two complementary ways. Based on their major element concentrations (Fe, Mg, O), carbon content, and abundance of chondrules, these meteorites naturally cluster into the distinct classes shown in Table 27. In addition, within each of these classes, the meteorites differ according to the degree that they have been thermally metamorphosed or experienced loss of volatile elements. This difference is referred to as the petrologic type (Table 28). For example, the Allende carbonaceous chondrite (Figure 66, top right) is classified CV3, indicating that it belongs to group CV (Table 27) and petrologic type 3 (Table 28).

Differences between chondrite types

Chondrites obviously differ from one another in several important respects. As has been pointed out, they vary in the extent to which they have undergone thermal metamorphism. Another important distinction is between the more abundant ordinary chondrites (of which there are three principal kinds) and the rarer chondritic meteorites that exhibit significant chemical differences. One of these types is the enstatite chondrite, which is, among other things, chemically more reduced than the ordinary chondrites. Almost all the iron in these meteorites, for example, is in metallic form. As a result, most of the abundant silicate mineral, pyroxene, is present as nearly pure enstatite ($MgSiO_3$) rather than in magnesium-iron solid solution minerals, such as bronzite and hypersthene, found in the ordinary chondrites. In enstatite chondrites, the readily oxidized element silicon is even found in the reduced state, and calcium occurs as the sulfide mineral oldhamite (CaS) rather than in its more usual silicate forms.

Other very important varieties of chondrites are grouped together as the carbonaceous chondrites. As their name implies, they characteristically contain more carbon (0.5 to 5 percent) than the ordinary chondrites (only about 0.1 percent). The mineral constituents of the carbonaceous chondrites are less chemically equilibrated with one another than even the unequilibrated ordinary chondrites. In many cases, one finds in the same meteorite carbonaceous material that formed at low temperatures and inclusions of the most refractory minerals—perovskite ($CaTiO_3$), hibonite ($CaAl_{12}O_{19}$), and melilite (solid solutions of $Ca_2Al_2SiO_7$ and $Ca_2MgSi_2O_7$).

CI carbonaceous chondrites

Perhaps the most interesting type of meteorite is the CI carbonaceous chondrite. Strictly speaking, one could legitimately question why such meteorites are called chondrites at all inasmuch as they do not contain chondrules. When compared with solar abundances (Figure 73), however, it turns out that they are the least differentiated meteorites of all, and in making a classification scheme it certainly makes sense to group them with the other undifferentiated meteorites. In accordance with the correlation already observed between chemical undifferentiation and

Table 28: Classification of Undifferentiated Meteorites in Terms of Petrologic Type and Metamorphic Grade*

	petrologic type					
	1	2	3	4	5	6
Homogeneity of olivine and pyroxene compositions	—	mean deviations of pyroxene ≥ 5%, of olivine ≥ 3%		5% > mean pyroxene deviation > 0%	uniform ferromagnesian minerals	
Structural state of low-Ca pyroxene	—	predominantly monoclinic		abundant monoclinic crystals	orthorhombic	
Degree of development of secondary feldspar	—	absent		predominantly as microcrystalline aggregates		clear, interstitial grains
Igneous glass	—	clear and isotropic primary glass, variable abundance		turbid if present	absent	
Metallic minerals	—	taenite absent or very minor (Ni < 200 mg/g)	kamacite and taenite (Ni > 200 mg/g) present			
Mean Ni content of sulfide minerals	—	> 5 mg/g	< 5 mg/g			
Overall texture	no chondrules	very sharply defined chondrules		well-defined chondrules	chondrules readily delineated	poorly defined chondrules
Texture of matrix	all fine-grained, opaque	much opaque matrix	opaque matrix	transparent microcrystalline matrix	recrystallized matrix	
Bulk carbon content	30–50 mg/g	8–26 mg/g	2–10 mg/g	< 2 mg/g		
Bulk water content	180–220 mg/g	20–160 mg/g	3–30 mg/g	< 15 mg/g		

*Following R. Van Schmus and J. Wood in *Geochim. Cosmochim. Acta* 31:747 (1967).
Source: J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

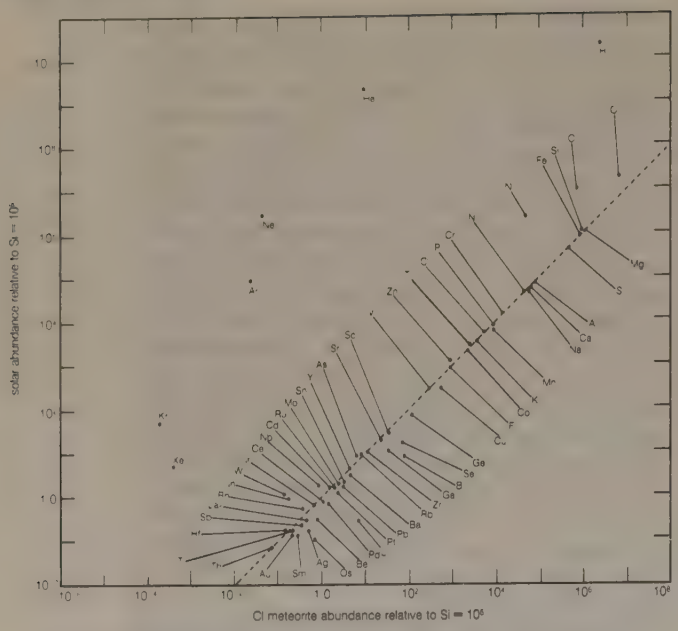


Figure 73: *Comparison of solar and meteoritic carbonaceous chondrite (CI) abundances of elements (relative to silicon = 10^6).*
For most elements, the solar and meteoritic abundances are very similar and therefore fall near a line with a slope of 45°. For the most volatile elements (hydrogen, noble gases, carbon, nitrogen, oxygen), the meteorites are highly depleted relative to the solar composition.

Based on data from *Meteorites: Their Record of Early Solar-System History* by John T. Wasson, W.H. Freeman and Company, 1985, and B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971), Gordon and Breach, Science Publishers, Inc.

chemical disequilibrium, the constituents of CI carbonaceous chondrites are far from equilibrium. The iron in these meteorites, for example, is highly oxidized, most of it occurring in the ferric iron-bearing mineral magnetite (Fe_3O_4). At the same time, carbon is present as highly reduced complex hydrocarbons. In equilibrium, the carbon would be oxidized to carbon monoxide and carbon dioxide, and the iron would be reduced to metallic iron.

Because CI chondrites are the most undifferentiated meteorites known, it has been speculated that, unlike most meteorites, they are of cometary rather than of asteroidal origin, since comets are believed to represent the most unaltered material in the solar system. There are difficulties in accepting this speculation as being correct. For example, detailed study of these meteorites shows that, in spite of their chemically undifferentiated and unequilibrated nature, they have had a complex chemical and physical history and are not simply a collection of interstellar dust. On the other hand, scientific knowledge about the nature and origin of comets is still limited, so that it would be unwise to dismiss this intriguing hypothesis prematurely.

Differentiated meteorites. Differentiated meteorites exhibit the type of chemical fractionation one would expect to occur on a planetary body that underwent core formation and magmatic differentiation similar to that observed in terrestrial and lunar volcanic rocks. Indeed, at one time it was thought likely that the most abundant type of differentiated stony meteorite, the basaltic achondrites, were actually lunar mare basalts. Their similarities to the lunar basalts subsequently returned by the Apollo missions showed that this was by no means a farfetched idea, but detailed considerations such as oxygen isotope ratios showed that it was not correct.

Iron meteorites and achondrites

In classifying differentiated meteorites, the major division is made between the iron meteorites and the stony differentiated meteorites, the achondrites ("not chondrites"). In addition, there are a number of stony-iron meteorites that contain mixtures of large masses of nickel-iron metal and differentiated silicate rock. The usual classification scheme for differentiated meteorites is given in Tables 29 and 30.

There are compelling reasons for believing that, like the chondrites, all differentiated meteorites, with the exception of those few from the Moon and possibly Mars, are asteroidal fragments. Because asteroids are too small to have experienced the accretional and long-lived radioactive heating that powers igneous processes on the inner planets, this may seem surprising. A partial solution to this enigma comes from radiometric dating, which shows that this differentiation took place very early in the history of the solar system, about 4.6 billion years ago. At that time other heating mechanisms may have been available—*e.g.*, heating by short-lived radioactive isotopes such as aluminum-26 (which has a half-life of about 700,000 years) or inductive heating by intense solar activity. It is even possible that asteroids from which differentiated meteorites seem to be derived are fragments of much larger differentiated planetesimals originally present in the inner planet region during the development of the Earth and Venus and which were stored in quasi-stable orbits in the innermost asteroid belt (2.2 AU) of the early solar system.

Specific asteroidal source regions for recovered meteorites. There is compelling, even if circumstantial, evidence that nearly all meteorites are derived from the asteroid belt. For scientific understanding of this matter to be complete, it is necessary to know which asteroids are the sources of particular types of meteorites and the mechanisms by which meteorites are transported from the asteroid belt to the Earth. It is possible that the ultimate answer will not be found until asteroids are explored by spacecraft. Nevertheless, considerable information relevant to this question is already available.

For the most abundant meteorite type, the ordinary

Table 29: Classification of Iron Meteorites*

class	subclass	symbol	nickel (%)	Widmanstätten structure kamacite (α-iron) bandwidth (mm)	percentage of total irons
Octahedrites		O	9–13		76
	coarsest	Ogg		>3.3	4
	coarse	Og		1.3–3.3	17
	medium	Om		0.5–1.3	40
	fine	Of		0.2–0.55	10
	finest	Off		<0.2	1
	plessitic	Opl		<0.2 (kamacite spindles)	4
Hexahedrites		H	5.5	pure kamacite	10
Nickel-rich ataxites		D	9–69	fine intergrowth of kamacite and taenite (γ iron; called plessite)	6

*Differentiated meteorites of this type consist principally of nickel-iron metal and iron sulfide; they make up 4 percent of all meteorite falls. The classification given here is structural. Another classification is based on the trace element concentrations of iron meteorites, particularly gallium, germanium, and iridium. This classification tends to group the same meteorites that are grouped in the structural classification, but the grouping is not exactly the same. For either classification, a significant number (10–15 percent) of meteorites do not fit well and must be classified "anomalous."
Sources: B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985); and V.F. Buchwald, *Handbook of Iron Meteorites* (1975).

Table 30: Silicate-Rich Differentiated Meteorites

class	principal silicate minerals	iron metal (weight %)	Fe-Mg silicate $\frac{FeO}{(FeO + MgO)}$	percentage of total observed falls
Pallasite	olivine	28–88	11–14	0.3
Mesosiderite	orthopyroxene plagioclase	30–55	23–27	0.8
Achondrites				
Basaltic-eucrite	pigeonite, plagioclase	<0.1	50–67	2.7
Basaltic-howardite	orthopyroxene, plagioclase	<1	25–40	2.4
Enstatite (aubrite)	enstatite	1	0	1
Ureilite	olivine, clinopyroxene	0.3–6	10–25	0.4
Diogenite	hypersthene	<1	25–27	1
Shergottite*	pigeonite, augite, plagioclase	0	59–76	0.3
Nakhlite*	augite, olivine	0	69	
Chassignite*	olivine	0	53	

*Isotopic and geochemical data suggest that these meteorites are from the planet Mars. Several achondrites found in Antarctica are of lunar origin. A small number (0.3 percent of total observed falls) do not fit into any of these categories and are classified as "anomalous."
Sources: B. Mason, *Handbook of Elemental Abundances in Meteorites* (1971); A.L. Graham, A.W.R. Bevan, and R. Hutchison, *Catalogue of Meteorites* (1985); and J.T. Wasson, *Meteorites, Their Record of Early Solar-System History* (1985).

Orbits of Earth-impacting ordinary chondrites

chondrites, there is a fairly large body of evidence that indicates that most of these meteorites strike the Earth while traveling in a rather special type of orbit. Such an orbit has its perihelion just inside the Earth's orbit, as well as low inclination (less than about 10°) and fairly high eccentricity (about 0.6). This has been determined on the basis of the time of day during which meteorites are observed to fall. It has been noted that about twice as many ordinary chondrites fall during the daylight afternoon hours as during daylight hours before noon. This fact requires that the meteorite must be moving faster than the Earth at the time of impact and, therefore, must be near its perihelion according to Keplerian dynamics. Quantitative consideration of the implications of time of fall data is entirely consistent with evidence obtained by visual observations of meteorite radiants and with about 30 fireballs identified by the Prairie Network as being, at least physically, equivalent to stony meteorites (see above).

This distribution of orbits places strong constraints on the source region in the asteroid belt from which ordinary chondrites can be derived. If meteorites in the kilogram mass range are to be derived directly from the asteroid belt, the mechanism by which they are extracted from the belt and transferred to an Earth-crossing orbit must operate quite rapidly on a time scale of a few million years. Only then can a meteoroid escape destruction by collision while near its aphelion in the asteroid belt. This collisional lifetime is not simply a theoretical result but is measured directly by the cosmic-ray exposure ages of meteorites, as discussed below.

High-energy galactic cosmic rays—primarily protons—have a range of penetration on the order of a few metres in meteoritic material. Any meteoroid of smaller dimensions will be radiated throughout by this proton bombardment. The high-energy protons cause spallation reactions (nuclear interactions that result in the release of many nucleons) on the abundant elements in the meteoroidal target. As a consequence, a large number of otherwise rare isotopic species, both stable and radioactive, are produced. These include the stable noble gas isotopes helium-3, neon-21, and argon-36, as well as the rare isotope potassium-40 and various short- and moderately long-lived radioactive isotopes, including hydrogen-3 (with a half-life of 12.26 years), beryllium-7 (53.29 days), beryllium-10 (1.6×10^6 years), aluminum-26 (7.2×10^5 years), manganese-53 (3.7×10^6 years), and cobalt-60 (5.272 years). The concentration of the shorter-lived radioactive isotopes can be used to monitor the cosmic-ray bombardment rate, and the accumulation of the stable species (*e.g.*, neon-21) measures the time in the past that this bombardment began—*i.e.*, the time that the meteoroidal fragment was separated from a larger object that was large enough to have shielded it from cosmic-ray bombardment. For chondritic meteorites, the distribution of cosmic-ray exposure age falls off quite quickly with age. This is to some extent a consequence of the dynamic evolution of the meteoroid orbits but for the most part should be attributed to a "collision half-life" of 5 to 10 million years.

Cosmic-ray exposure age

There are only two processes known that can accelerate meteoroidal fragments into Earth-crossing orbits on this short time scale given by cosmic-ray exposure ages. These processes are direct collisional ejection at velocities of about five kilometres per second and gravitational acceleration by dynamic resonances in the asteroid belt, of which the 3:1 resonance at 2.5 AU is of dominant importance. In a hypervelocity collision, some material is ejected at the required high velocity, but the quantity of this material is small and most of it is pulverized by the associated shock pressures. High-velocity ejection is likely to be responsible for the occurrence of meteorites from Mars or the Moon, but it completely fails to provide the observed quantity of meteoroids from the asteroid belt. The resonant mechanisms are therefore of much greater importance. Bodies orbiting the Sun with a semimajor axis near 2.5 AU will complete three revolutions about the Sun in the time that Jupiter, a strong source of gravitational perturbations, executes one revolution. The resulting resonant acceleration will cause the orbit of the body to become "chaotic," and its perihelion will become Earth-crossing in about one million years.

The calculated quantity of asteroidal material in the meteorite size range delivered to the Earth from the 3:1 resonance agrees well with the 10^8 gram per year terrestrial impact rate of ordinary chondritic meteoroids. Moreover, this resonant acceleration presents a natural mechanism for concentrating meteoroid perihelia near the Earth's orbit, thereby explaining the special distribution of orbits observed for ordinary chondrites. In fact, as it turns out, meteoroids derived by this resonance acceleration mechanism should be overly concentrated near 1 AU; the ratio of afternoon to morning falls should be about 3:1 instead of the observed 2:1. This discrepancy is removed when one takes into account the fact that larger asteroidal fragments (those reaching the size of Apollo objects) will also be accelerated into an Earth-crossing orbit by the same resonant mechanism. Meteorite-size fragments will be produced as collision debris from these larger bodies, but they will not have the special distribution of orbits exhibited by the smaller asteroidal fragments introduced more directly into Earth-crossing orbits. When the contribution from resonant asteroidal ejecta over the entire size range is averaged, the predicted and observed orbital distribution match rather well.

Orbital statistics are not very well known for other meteorite types. It does appear that the differentiated basaltic achondrites fail to share the special orbital distribution observed for the ordinary chondrites. Basaltic achondrites are most likely the collision debris of Apollo objects that were extracted by another resonance mechanism known to operate in the innermost belt; this mechanism can be shown to be effective only in providing larger Earth-crossing bodies. Although more work remains to be done on

this problem, it seems likely that known resonant mechanisms are adequate to explain the dynamical processes by which all classes of meteorites are delivered from their asteroidal source regions into Earth-crossing orbits.

Derivation of meteoritic material from these designated regions of the inner asteroid belt implies that asteroids in such regions have the chemical and mineralogical composition observed in the meteorites. The surface mineralogical composition of asteroids can be determined directly by Earth-based reflectance spectrometry. These measurements have been made for most of the larger asteroids. Although no two reflectance spectra are exactly alike in detail, most asteroids fall into one of two general groups, the S class and the C class. The S-class asteroids have moderate albedos (though comparatively high overall) and contain mixtures of olivine (magnesium, iron silicates), pyroxene (silicates containing magnesium, iron, calcium, and aluminum), and metallic iron. These are the same minerals found in ordinary chondrites and in basaltic achondrites. The C-class asteroids have low reflectance, and their more featureless spectra indicate the presence of light-absorbing opaque minerals. It is plausible to consider these asteroids as candidate sources for carbonaceous meteorites.

Chemical and mineralogical similarities between the major classes of asteroids and meteorites

When considered in more detail, there are certain difficulties in identifying the required number of S-class asteroids as ordinary chondrite sources. Although qualitatively the mineralogy of these asteroids agrees with that of ordinary chondrites, the proportions of the constituent minerals to one another does not match as well. In particular, the S-class asteroids appear to have, at least on their surfaces, about twice as much metallic iron as the ordinary chondrites. The solution to this discrepancy is not known at present. It may be that the surfaces of the asteroids are not truly representative of their interiors, having been altered by collisional bombardment or exposure to solar radiation and particles. Another possibility is that a systematic change in mineralogical composition occurs during the process of collisional grinding of asteroids into ever-smaller bodies, ultimately to those of meteorite size. There is some indication that the discrepancy is less serious for small asteroidal fragments observed as Apollo objects, which is consistent with this hypothesis.

Meteorites, asteroids, and the early solar system. A prime long-range objective of exploring the Moon and planets by spacecraft is to collect samples of these bodies for detailed laboratory study. With the few exceptions noted earlier, meteorites are samples of asteroids delivered to Earth by fairly well-understood natural mechanisms. It also is believed that samples collected from asteroids will prove especially valuable, because such small bodies are most likely to be "primitive" and thus retain the record of events that occurred during their own formation and that of the solar system in general. The study of meteorites is already providing scientists with much valuable information directly related to this matter.

Isotopic records. Meteorites indicate to investigators that the asteroid belt must always have been a relatively tranquil region of the solar system. Some unequilibrated chondrites have inherited and preserved without complete mixing the remnants of presolar interstellar grains. This is demonstrated by variations of the ratios of the oxygen isotopes oxygen-16, oxygen-17, and oxygen-18 within a single meteorite and between different meteorites. It is well known that the oxygen isotopes are fractionated by natural chemical processes. Variations in the ratio of oxygen-18 to oxygen-16, for example, form the basis for paleotemperature studies of ancient terrestrial sedimentary rocks by making use of the temperature dependence of the isotopic chemical equilibrium between seawater and the calcium carbonate that forms the shells of marine organisms. It is characteristic of this natural fractionation dependent on isotopic mass that the degree of fractionation is proportional to the difference of the masses of the isotopes. Thus, in these marine sediments, the variation in the ratio oxygen-18:oxygen-16 is twice that of oxygen-17 to oxygen-16. In some unequilibrated chondrites, however, the variations in the oxygen-18 to oxygen-16 ratios are equal to those in oxygen-17:oxygen-16. It is possible to devise special chemical processes that could produce fractionation of this kind, but it is much more likely that the variations found in the meteorites are caused by nuclear processes that predated the formation of the Sun and solar system. The interstellar grains that were the carriers of these isotope anomalies were probably formed in stellar atmospheres and preserved the signatures of the isotope-formation processes of stars of particular types. These isotopic variations of nuclear origin are found not only in oxygen but also in other less abundant elements, including neon, xenon, titanium, and chromium.

Remnants of presolar interstellar grains

Probable early evolution of the solar system based on meteoritic evidence. The need to provide an environment sufficiently tranquil to preserve this isotopic record, as well as other fragile relics of early solar system events observed in meteorites, places important constraints on the formation of the solar system. If one examines the distribution of matter in the present solar system, it is seen that the density is high both in the region of the inner planets and in the region of the giant planets in the outer solar system but very low in the wide space between Mars and Jupiter. This fact itself is surprising: Why should the solar nebula from which the solar system formed have had a great hole in it? The answer is that it probably did not. In order for asteroids to have formed and developed at all on the time scale of a few million years indicated by the radiometric dating of meteorites, densities more like those in the regions occupied by the giant planets would have been required, as shown by theoretical calculations. It is difficult to escape the conclusion that the quantity of matter in what is now the asteroid belt must have been much greater, perhaps by as much as 10,000 times the quantity observed there today. Therefore some natural process has to have removed almost all the material in this region of the solar system after the formation of asteroidal bodies, but in a sufficiently gentle way to have preserved the relics of presolar and early solar events found in meteorites.

Although the details are not yet understood, it seems most likely that the formation of the giant planets, particularly Jupiter, quickly resulted in the evacuation of most of the matter from this region of the solar system. This means that Jupiter formed rapidly, before bodies in the asteroid belt had grown to become full-fledged planets. The mineralogical and chemical record of the undifferentiated meteorites is not compatible with their once having been part of a planet even as large as the Moon. Also, the very energetic collisional events that would be associated with the dispersal of large planets in the asteroid belt would preclude preservation of the observed relics.

These constraints, for the most part based on meteoritic evidence, define a conceivable chain of events for the early evolution of the solar system beyond the region of the inner planets. Even though the density of matter was at least as great in the asteroid belt as in the region of Jupiter, planetary growth must have been more rapid in Jupiter's vicinity than in the asteroid belt. (Some suggestions have been made as to how this may be possible, but it remains to be seen if they are satisfactory.) Within about one million years, proto-Jupiter(s) began to capture the massive quantities of hydrogen and helium from the solar nebula that constitute most of the giant planet today. At the same time, thousands of large asteroids greater than 100 kilometres in diameter, including some as big as the largest present-day asteroids but not much bigger, had formed. Shortly thereafter, as Jupiter approached its present mass, most of the residual nebular gas was removed by the intense solar ultraviolet radiation characteristic of young stars.

Planetary growth and changes in mass distribution in the outer solar system

Up until the time Jupiter approached its present mass, the asteroids moved in nearly circular orbits in accordance with the weak mutual gravitational perturbations expected for small bodies. During the final formation of Jupiter and Saturn and the removal of the nebular gas, the changing mass distribution in the outer solar system caused waves of resonant perturbations to sweep through the asteroid belt, increasing the eccentricities and inclinations of the asteroids to the moderate values observed today. Because the asteroids had not grown larger than the asteroids of today, collisions at the relative velocities of about five kilometres per second, associated with their present-day

eccentricities and inclinations, would begin to grind down the smaller bodies. This would occur without the shock effects associated with disruption of larger planetary bodies at higher impact velocities.

The foregoing scenario of planetary growth is certain to be wrong in detail and may well be wrong altogether. Nevertheless, without the samples of asteroids provided by recovered meteorites, there would be little observational basis at all for formulating models of this kind. All the qualitative statements made above should be considered not as established facts but rather as statements of theoretical problems that need to be thoroughly worked out. Their results must be compared with old and new observational data and reformulated and reworked until a satisfactory understanding of planetary formation is achieved.

Thermal evolution of the solar nebula and planetesimals. The available meteoritic evidence is relevant to many other unsolved central questions concerning the early solar system, including some that bear on the way the Sun was formed and its early history. One of these is the question of the thermal evolution of the solar nebula and the planetesimals that formed and grew within it.

As discussed above, the environment of the early asteroid belt must have been a rather "tranquil" one. By tranquil, one actually means thermally, rather than physically, uneventful. The collisions that led to the removal of most of the material from the asteroid belt were highly disruptive, and there is ample evidence of extensive physical disruption in the textures of even the most primitive meteorites. Yet, the preservation of primordial relics, such as the isotope anomalies, argues against widespread heating of the asteroidal region to temperatures as high as 1,000 K.

Low-temperature origin of the asteroidal region

A relatively cool solar nebula at distances this far from the Sun is in agreement with most current theoretical calculations of the formation of the Sun. In addition to the evidence for an overall low-temperature origin for this region of the solar system, however, the meteoritic record clearly shows the imprint of high-temperature events of major importance, which are not well understood at present. The most apparent of these is the very abundant presence of chondrules in all undifferentiated meteorites except the CI chondrites. Chondrules appear to have once been melted droplets primarily of silicate composition and would have required temperatures of about 1,500 K to have formed.

Formation of chondrules

If chondrules were a relatively rare meteoritic curiosity, one could legitimately consider them an interesting detail to be explained someday but not a matter of central importance. Yet, the fact that chondrules (or their broken fragments) make up most of the mass of the most abundant class of meteorites, the ordinary chondrites, and a major portion of other chondrites, indicates that their formation must have been of major importance in the early solar system. Even if ordinary chondrites formed only within a restricted region of the asteroid belt adjacent to the 3:1 Kirkwood gap (*i.e.,* between 2.44 and 2.56 AU), this is still about 10 percent of the entire asteroid belt. It also is likely that other chondrule-bearing meteorites formed outside this region, even though their asteroidal sources may be in that region today.

It seems impossible that chondrules are the product of igneous differentiation because they are of nearly undifferentiated solar composition except for the most volatile elements. Nor does it seem likely that they are impact droplets, such as those found in the lunar soil. The expected low impact velocities of asteroidal planetesimals argue against a ubiquitous high-velocity impact environment. Thus, there seems to be no way by which the chondrules could have formed in or on planetesimals. In all likelihood, they were formed in the solar nebula. At the same time, the evidence for rapid cooling of chondrules argues against their formation by large-scale condensation from a very hot solar nebula. Local, transient heating events appear to have been important on a wide scale in the solar nebula, but the nature and cause of these events remain unknown.

A problem of similar difficulty is that of the origin of the large (up to more than one centimetre in diameter), highly refractory inclusions found in the CV meteorites (see Table 27), especially prominent in the Allende carbonaceous chondrites. Though not as ubiquitous as chondrules, these inclusions are by no means of negligible abundance. Unlike chondrules, they are highly fractionated chemically, apparently as a result of more prolonged heating to about 1,500 K. Because many of the isotopic anomalies are associated with these refractory inclusions, interpretation of this important evidence is limited by a poor understanding of their origin.

Evidence of internal heating in differentiated meteorites

The foregoing thermal events most likely occurred in the solar nebula rather than on growing asteroidal bodies. In addition, thermal effects are observed in meteorites associated with internal heating. The most apparent of these are the differentiated meteorites, which probably represent about 10 percent of the asteroidal region sampled. The asteroids from which these meteorites were fragmented, though probably formed from a relatively cool solar nebula, experienced internal heating, core formation, and igneous differentiation within a few million years of the formation of the solar nebula itself. Clear and detailed evidence for this is based on (1) radiometric dating of the minerals formed by this igneous differentiation, (2) the mineral assemblages of the resulting igneous rocks, and (3) the slow cooling that produced large crystals of differentiated minerals (*e.g.,* the so-called Widmanstätten structure observed in many nickel-iron meteorites). The actual process responsible for this heating is yet unknown, but several good possibilities are being evaluated. These include heating by the short-lived radioactive isotope aluminum-26, heating by electric currents induced by early solar activity, and accretional heating of planetesimals in the terrestrial planetary region, followed by fragmentation and transfer to the innermost asteroid belt.

The ordinary chondrites also experienced heating after the formation of chondritic planetesimals, but not enough to produce melting. As a result, chondritic material, presumably once resembling the unequilibrated ordinary chondrites, was metamorphosed to produce the more abundant equilibrated ordinary chondrites. The time scale for this metamorphism is, within uncertainties, the same as that which produced the parent asteroids of the differentiated meteorites. It is plausible that some of the same heat sources (*e.g.,* aluminum-26) may have been responsible.

(G.W.We.)

BIBLIOGRAPHY

General works. Good overviews of the solar system include J. KELLY BEATTY, CAROLYN COLLINS PETERSON, and ANDREW CHAIKIN (eds.), *The New Solar System,* 4th ed. (1999); and DAVID MORRISON and TOBIAS OWEN, *The Planetary System,* 2nd ed. (1996). Individual solar system objects are treated in the excellent series of books published by the University of Arizona Press: FAITH VILAS, CLARK R. CHAPMAN, and MILDRED SHAPLEY MATTHEWS (eds.), *Mercury* (1988); RICHARD P. BINZEL, TOM GEHRELS, and MILDRED SHAPLEY MATTHEWS (eds.), *Asteroids II* (1989); JAY T. BERGSTRALH, ELLIS D. MINER, and MILDRED SHAPLEY MATTHEWS (eds.), *Uranus* (1991); HUGH H. KEIFFER *et al.* (eds.), *Mars* (1992); DALE P. CRUIKSHANK (ed.), *Neptune and Triton* (1995); S.W. BOUGHER, D.M. HUNTEN, and R.J. PHILLIPS (eds.), *Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment* (1997); and S. ALAN STERN and DAVID J. THOLEN (eds.), *Pluto and Charon* (1997). Annually revised orbital and physical data about planets, moons, and selected comets and asteroids appear in *The Astronomical Almanac,* published by the U.S. Naval Observatory *et al.*; *The Observer's Handbook,* published annually by the Royal Astronomical Society of Canada, provides excellent information for observing solar system objects with the naked eye or small telescopes. International reports of research on solar system objects appear regularly in *The Astrophysical Journal,* published by the American Astronomical Society and University of Chicago; *The Astronomical Journal,* published by the American Institute of Physics and American Astronomical Society; *Astronomy and Astrophysics,* published by the European Southern Observatory; *Icarus,* a journal of solar system studies published by the American Astronomical Society; *Journal of Geophysical Research,* published by the American Geophysical Union; and *Annual Review of Earth and Planetary Sciences.* A superior monthly periodical for the nonprofessional, with regular coverage of the solar system and its constituents, is *Sky and Telescope.*

(T.C.O.)

Origin of the solar system. An excellent collection of papers on the general subject of solar system origin appears in VINCENT MANNINGS, ALAN P. BOSS, and SARA S. RUSSELL (eds.), *Protostars & Planets IV* (2000); the volume includes papers about newly discovered planets around other stars. Volumes of original technical

articles by different authors on facets of the topic are RICHARD GREENBERG, ANDRÉ BRAHIC, and MILDRED SHAPLEY MATTHEWS (eds.), *Planetary Rings* (1984); JOHN F. KERRIDGE and MILDRED SHAPLEY MATTHEWS (eds.), *Meteorites and the Early Solar System* (1988); and S.K. ATREYA, J.B. POLLACK, and MILDRED SHAPLEY MATTHEWS (eds.), *Origin and Evolution of Planetary and Satellite Atmospheres* (1989). The formation of the inner planets has been extensively studied by GEORGE W. WETHERILL, "Formation of the Earth," *Annual Review of Earth and Planetary Sciences*, 18:205–256 (1990), a review of his and others' work in the field. (T.C.O.)

The Sun. Popular works on the Sun include HERBERT FRIEDMAN, *Sun and Earth* (1986); RONALD G. GIOVANELLI, *Secrets of the Sun* (1984); and ROBERT W. NOYES, *The Sun, Our Star* (1982). KARL HUFBAUER, *Exploring the Sun: Solar Science Since Galileo* (1991), chronicles the history of developments in this field. Works of a more technical nature include PETER FOUKAL, *Solar Astrophysics* (1990); R.O. PEPIN, J.A. EDDY, and R.B. MERRILL, *The Ancient Sun: Fossil Record in the Earth, Moon, and Meteorites* (1980); MICHAEL STIX, *The Sun: An Introduction* (1989), showing the many techniques and ideas utilized to study the Sun; WASABURO UNNO *et al.*, *Nonradial Oscillations of Stars* (1979); HAROLD ZIRIN, *Astrophysics of the Sun* (1988); A.N. COX, W.C. LIVINGSTON, and M.S. MATTHEWS (eds.), *Solar Interior and Atmosphere* (1991); and papers from three Skylab Solar Workshops: JACK B. ZIRKER (ed.), *Coronal Holes and High Speed Wind Streams* (1977); PETER A. STURROCK, *Solar Flares* (1980); and FRANK Q. ORRALL, *Solar Active Regions* (1981). (H.Zi.)

The planets and their moons. *Mercury:* ROBERT G. STROM and ANN L. SPRAGUE, *Exploring Mercury: The Iron Planet* (2003), by two expert researchers in the field, was written during the development of the Messenger spacecraft mission. J. KELLY BEATTY, CAROLYN COLLINS PETERSEN, and ANDREW CHAIKIN (eds.), *The New Solar System*, 4th ed. (1999), contains several chapters of information related to Mercury, especially in the context of the terrestrial planets. FAITH VILAS, CLARK R. CHAPMAN, and MILDRED SHAPLEY MATTHEWS (eds.), *Mercury* (1988), is a comprehensive collection of technical chapters and overviews covering knowledge of the planet to the time of publication, with an extensive bibliography. Nontechnical treatments of the Mariner 10 mission are found in BRUCE MURRAY and ERIC BURGESS, *Flight to Mercury* (1977); and JAMES A. DUNNE and ERIC BURGESS, *The Voyage of Mariner 10: Mission to Venus and Mercury*, NASA Special Publication 424 (1978); the latter is a well-illustrated account of the drama surrounding the mission, with pictures of people, instruments, and trajectories plus a large collection of images and maps of Mercury. An excellent collection of Mariner 10 photographs is available in MERTON E. DAVIES *et al.* (eds.), *Atlas of Mercury* (1976). (C.R.C.)

Venus: LADISLAV E. ROTH and STEPHEN D. WALL (eds.), *The Face of Venus: The Magellan Radar-Mapping Mission* (1995), is a good post-Magellan popular-level book on the Venusian surface, with many excellent illustrations. Still informative pre-Magellan popular-level treatments include GARRY E. HUNT and PATRICK MOORE, *The Planet Venus* (1982); and ERIC BURGESS, *Venus, an Errant Twin* (1985). The scientific understanding of Venus is definitively and comprehensively summarized in S.W. BOUGHER, D.M. HUNTEN, and R.J. PHILLIPS (eds.), *Venus II* (1997), a collection of papers written after the Magellan and Galileo missions. MIKHAIL YA. MAROV and DAVID H. GRINSPOON, *The Planet Venus* (1998), provides an excellent post-Magellan treatment of Venusian geology and most other aspects of the planet. A pre-Magellan overview of Venus' surface is given by ALEXANDER T. BASILEVSKY and JAMES W. HEAD III, "The Geology of Venus," *Annual Review of Earth and Planetary Sciences*, 16:295–317 (1988). An interesting collection of papers by Soviet scientists in English on Venus is V.L. BARSUKOV *et al.* (eds.), *Venus Geology, Geochemistry, and Geophysics* (1992). The results of major spacecraft missions to Venus are reported in several journals: on Pioneer Venus, *Journal of Geophysical Research*, 85:7573–8337 (1980); V.L. BARSUKOV *et al.*, "The Geology and Geomorphology of the Venus Surface as Revealed by the Radar Images Obtained by Veneras 15 and 16," *Journal of Geophysical Research*, part B, *Solid Earth and Planets*, 91(B4):D378–D398 (March 30, 1986); *Science*, 253:1457–1612 (Sept. 27, 1991), an issue devoted to the Galileo flyby of Venus; and two issues of *Journal of Geophysical Research*, part E, *Planets*, vol. 97, devoted to detailed Magellan mission results, no. 8 (Aug. 25, 1992) and no. 10 (Oct. 25, 1992). (S.W.S.)

Earth: PETER D. WARD and DON BROWNLEE, *Rare Earth: Why Complex Life Is Uncommon in the Universe*, 2nd rev. ed. (2003), lays out for general readers a case for the uncommon nature of planet Earth. JONATHAN I. LUNINE, *Earth: Evolution of a Habitable World* (1999); and JOHN J.W. ROGERS, *A History of the Earth* (1993), provide comprehensive introductions to Earth from, respectively, planetological and geologic perspectives. WILLIAM K. HARTMANN, *The History of Earth: An Illustrated Chronicle of an Evolving Planet* (1991), is a lavishly illustrated introduction. A graduate-level text comprising excellent chapters on its subject matter is ROBIN M. CANUP and KEVIN RIGHTER (eds.), *Origin of the Earth and Moon* (2000). A popular and balanced account of the physical causes of mass extinctions is given in CHARLES FRANKEL, *The End of the Dinosaurs: Chicxulub Crater and Mass Extinctions* (1999; originally published in French, 1996). The periodical *Scientific American* (monthly) is an excellent and accessible source for the latest thinking on Earth and planetary processes; see, for example, a good description of plate tectonics in IAN W.D. DALZIEL, "Earth Before Pangea," 272(1):58–63 (January 1995). The *State of the World* report (annual), published by the Worldwatch Institute, provides authoritative updates on Earth's vital signs for the general reader. (J.I.L.)

The Moon: Revival of interest in the Moon and in lunar exploration after the long post-Apollo hiatus is reflected in PAUL D. SPUDIS, *The Once and Future Moon* (1996), a readable and well-illustrated summary of recent scientific knowledge as well as arguments advocating continued exploration and eventual human return to the Moon. The rich and expanding scientific literature about the Moon is well represented in *Proceedings of Lunar and Planetary Science*, published collections of papers from the annual Lunar and Planetary Science Conference. The finest book about lunar exploration remains DAVIS THOMAS (ed.), *Moon: Man's Greatest Adventure*, rev. ed. (1973), splendidly illustrated, with text by scholars describing the history and culture, engineering and projects, and early scientific results of the great human drive that began in ancient times and culminated in the Apollo missions. A more modest but still comprehensive and well-illustrated book is PATRICK MOORE, *The Moon* (1981, reprinted 1984). DON E. WILHELMS, *The Geologic History of the Moon* (1987), is illustrated with many beautiful explanatory pictures and drawings. Examples of relevant literature on theories of lunar origin include W.K. HARTMANN, R.J. PHILLIPS, and G.J. TAYLOR (eds.), *Origin of the Moon* (1986); and A.V. VITIAZEV, G.V. PECHERNIKOVA, and V.S. SAFRONOV, *Planety zemnoy gruppy: proiskhozhdeniye i rannyaya evolyutsiya* ("Terrestrial Planets: Their Origin and Early Evolution"; 1990). STUART ROSS TAYLOR, *Planetary Science: A Lunar Perspective* (1982); and W.W. MENDELL (ed.), *Lunar Bases and Space Activities of the 21st Century* (1985), are collections of specialized papers, although most of the material is accessible to the general reader. The realization that the Moon is an enormous storehouse of resources that may be useful to humankind in the future has prompted several publications, including GERARD K. O'NEILL, *The High Frontier*, 3rd ed. (2000); a series of conference papers, *Space Manufacturing* (biennial); and a book commissioned by NASA to serve as a primary reference, GRANT HEIKEN, DAVID VANIMAN, and BEVAN M. FRENCH (eds.), *Lunar Sourcebook: A User's Guide to the Moon* (1991), which with its compendious bibliography is an excellent summary of what is known about the Moon. PETER ECKART (ed.), *The Lunar Base Handbook: An Introduction to Lunar Base Design, Development, and Operations* (1999), gives up-to-date information on lunar resources and their uses, plus engineering and scientific aspects of the creation of human habitats on the Moon. (J.D.Bu.)

Mars: DONALD GOLDSMITH, *The Hunt for Life on Mars* (1997), describes the events that led to the controversial claims for evidence of life in the Martian meteorite ALH84001. JOHN NOBLE WILFORD, *Mars Beckons: The Mysteries, the Challenges, the Expectations of Our Next Great Adventure in Space* (1990), provides a general introduction to the planet. All aspects of Mars science are addressed in HUGH H. KIEFFER *et al.* (eds.), *Mars* (1992), a huge, comprehensive technical summary of existing knowledge. VICTOR R. BAKER, *The Channels of Mars* (1982), is a well-illustrated discussion of how these features came about and their implications for the history of Mars. MICHAEL H. CARR, *The Surface of Mars* (1981), is a profusely illustrated survey of scientists' perception of Mars after the Viking missions, and *Water on Mars* (1996) discusses the role that water has played in the evolution of the Martian surface. Volumes of papers detailing the results of Mars missions appear in the *Journal of Geophysical Research:* for Viking, 82, no. 28 (1977); for Mars Pathfinder, 104, no. E4 (1999), and for Mars Global Surveyor, 106, no. E4 (2001). These results are summarized in several excellent papers in *Nature* 412:207–253 (July 12, 2001). (M.C.Ma./M.H.Ca.)

Jupiter: A comprehensive, popular-level review of knowledge of the Jovian system, including early results from the Galileo space probe, is RETA BEEBE, *Jupiter: The Giant Planet*, 2nd ed. (1997). Detailed descriptions of the Galileo mission and its findings are provided in DAVID M. HARLAND, *Jupiter Odyssey: The Story of NASA's Galileo Mission* (2000); and DANIEL FISCHER, *Mission Jupiter: The Spectacular Journey of the Galileo Spacecraft* (2001). The original reference for visual observations of Jupiter with telescopes of moderate size is BERTRAND M. PEEK, *The Planet Jupiter*, rev. by PATRICK MOORE (1981). A more recent account that includes results from the Voyager space probes is

JOHN H. ROGERS, *The Giant Planet Jupiter* (1995). Details regarding the Voyager missions may be found in DAVID MORRISON and JANE SAMZ, *Voyage to Jupiter* (1980). (T.C.O.)

Saturn: DAVID MORRISON, *Voyages to Saturn* (1982), is a nontechnical presentation. More advanced treatments are contained in TOM GEHRELS and MILDRED SHAPLEY MATTHEWS (eds.), *Saturn* (1984), a collection of essays. Journal articles include G.F. LINDAL, D.N. SWEETNAM, and V.R. ESHLEMAN, "The Atmosphere of Saturn: An Analysis of the Voyager Radio Occultation Measurements," *The Astronomical Journal,* 90(6):1136–1146 (June 1984), specific experimental results on the structure of Saturn's atmosphere; and two issues of *Journal of Geophysical Research,* part A, *Space Physics:* vol. 85, no. A11 (Nov. 1, 1980), devoted to Pioneer 11 results, and vol. 88, no. A11 (Nov. 1, 1983), devoted to Voyager 1 and 2 results. (W.B.H.)

Uranus: A.P. INGERSOLL, "Uranus," *Scientific American,* 256(1):30–37 (January 1987), provides an introductory review article with illustrations and diagrams. GARRY HUNT and PATRICK MOORE, *Atlas of Uranus* (1989), offers an in-depth introduction. The first reports of the Voyager 2 encounter written by the Voyager scientists are included in a set of 12 articles in *Science,* 233(4759):39–109 (July 4, 1986), which contains most of the best images of the planet, its satellites, and its rings. JAY T. BERGSTRALH, ELLIS D. MINER, and MILDRED SHAPLEY MATTHEWS (eds.), *Uranus* (1991), with chapters written by specialists in the field, is the definitive reference work on the subject. ELLIS D. MINER, *Uranus: The Planet, Rings, and Satellites* (1990), reviews current knowledge of the planet, with much background information on the Voyager mission. GARRY HUNT (ed.), *Uranus and the Outer Planets* (1982), is a pre-Voyager summary dealing almost exclusively with Uranus, with some interesting historical chapters. ERIC BURGESS, *Uranus and Neptune: The Distant Giants* (1988), also focuses primarily on Uranus. MARK LITTMANN, *Planets Beyond: Discovering the Outer Solar System,* updated and rev. ed. (1990), chronicles the history of the discovery of Uranus, Neptune, and Pluto. (A.P.I.)

Neptune: PATRICK MOORE, *The Planet Neptune* (1988), provides a good summary of pre-Voyager knowledge of the planet. A set of articles in *Science,* 246(4936):1417–1501 (Dec. 15, 1989), comprises the initial report of the Voyager findings at Neptune; more detailed reports of the findings are in a set of articles in *Journal of Geophysical Research, Supplement,* 96:18,-903–19,268 (Oct. 30, 1991); and ERIC BURGESS, *Far Encounter: The Neptune System* (1991), a popular work which also contains some discussion of Pluto. See also the work by Littmann cited above for Uranus. (E.D.M.)

Pluto: S.A. STERN and J. MITTON, *Pluto and Charon: Ice Worlds on the Ragged Edge of the Solar System* (1998, reissued 2000), is a popular-level discussion of the Pluto-Charon system. An authoritative collection of papers by experts in the field is presented in S.A. STERN and DAVID J. THOLEN (eds.), *Pluto and Charon* (1997). CLYDE W. TOMBAUGH and PATRICK MOORE, *Out of the Darkness, the Planet Pluto* (1980), is the story of the planet's discovery cowritten by its discoverer (Tombaugh). The detection of Charon is reported in articles by its discoverers in J.W. CHRISTY and R.S. HARRINGTON, "The Satellite of Pluto," *The Astronomical Journal,* 83(8):1005–08 (August 1978), and "The Discovery and Orbit of Charon," *Icarus,* 44(1):38–40 (October 1980). See also the work by Littmann cited above for Uranus. (T.C.O.)

Other constituents of the solar system. A.H. DELSEMME (ed.), *Comets, Asteroids, Meteorites: Interrelations, Evolution, and Origins* (1977), an extensive collection of colloquium papers, provides a scholarly overview of the minor solar system bodies addressed below. (A.H.De.)

Asteroids: Summary articles can be found in RICHARD P. BINZEL, M. ANTONIETTA BARUCCI, and MARCELLO FULCHIGNONI, "The Origins of the Asteroids," *Scientific American,* 265(4):88–94 (October 1991). Review and research papers are collected in RICHARD P. BINZEL, TOM GEHRELS, and MILDRED SHAPLEY MATTHEWS (eds.), *Asteroids II* (1989); and C.-I. LAGERKVIST *et al.* (eds.), *Asteroids, Comets, Meteors III* (1990). (E.F.T.)

Comets: General introductory works are ARMAND H. DELSEMME, "Whence Come Comets?" *Sky and Telescope,* 77(3):260–264 (March 1989), an elementary discussion on their origin; FRED L. WHIPPLE and DANIEL W.E. GREEN, *The Mystery of Comets* (1985); and ROBERT D. CHAPMAN and JOHN C. BRANDT, *The Comet Book: A Guide for the Return of Halley's Comet* (1984), a historical treatment. More advanced are R.L. NEWBURN, JR., M. NEUGEBAUER, and J. RAHE (eds.), *Comets in the Post-Halley Era,* 2 vol. (1991), containing reviews, summaries, and scientific papers by about 100 authors; K.S. KRISHNA SWAMY, *Physics of Comets* (1986), including a section on spectroscopy; and JOHN C. BRANDT and ROBERT D. CHAPMAN, *Introduction to Comets* (1981). DONALD K. YEOMANS, *Comets: A Chronological History of Observation, Science, Myth, and Folklore* (1991), is a comprehensive reference book on all cometary apparitions, full of anecdotes. LAUREL L. WILKENING and MILDRED SHAPLEY MATTHEWS (eds.), *Comets* (1982), is a definitive collection of essays covering all aspects at a technical level. See also D.A. MENDIS, H.L.F. HOUPIS, and M.L. MARCONI, "The Physics of Comets," *Fundamentals of Cosmic Physics,* 10 (1–4):1–380 (1985). BRIAN G. MARSDEN, *Catalogue of Cometary Orbits,* 7th ed. (1992), covers 1,353 cometary orbits, with detailed references and notes; a complementary volume is GARY W. KRONK, *Comets: A Descriptive Catalog* (1984). (A.H.De.)

Meteoroids, meteors, and meteorites: Introductory information can be found in HARRY Y. MCSWEEN, JR., *Meteorites and Their Parent Planets* (1987); ROBERT T. DODD, *Thunderstones and Shooting Stars: The Meaning of Meteorites* (1986); JOHN G. BURKE, *Cosmic Debris: Meteorites in History* (1986); ROBERT HUTCHISON, *The Search for Our Beginning: An Enquiry, Based on Meteorite Research, into the Origin of Our Planet and of Life* (1983); and JOHN A. WOOD, *Meteorites and the Origin of Planets* (1968). More advanced treatments are JOHN T. WASSON, *Meteorites: Their Record of Early Solar-System History* (1985), and *Meteorites: Classification and Properties* (1974); V.A. BRONSHTEN, *Physics of Meteoric Phenomena* (1983; originally published in Russian, 1981); and ROBERT T. DODD, *Meteorites: A Petrologic-Chemical Synthesis* (1981). A descriptive and historical treatment of iron meteorites, including beautiful photographs, is VAGN F. BUCHWALD, *Handbook of Iron Meteorites, Their Distribution, Composition, and Structure,* 3 vol. (1975). H.H. NININGER, *Out of the Sky: An Introduction to Meteorites* (1952, reprinted 1959), provides firsthand experiences of fall phenomena on a nontechnical level. See also D.E. BROWNLEE, "Cosmic Dust: Collection and Research," *Annual Reviews of Earth and Planetary Sciences,* 13:147–173 (1985). A catalog of known meteorites, including data regarding their fall, is A.L. GRAHAM, A.W.R. BEVAN, and R. HUTCHISON (eds.), *Catalogue of Meteorites,* 4th ed. rev. and enlarged (1985). There are two journals devoted to papers on meteorites and related bodies: *Meteoritika* (annual), published in Russia; and *Meteoritics and Planetary Science* (monthly). Many papers on meteorites are published in *Geochimica et Cosmochimica Acta* (semimonthly). (G.W.We.)

Sound

One question that has long been argued over, by philosophers as well as children, is whether a tree falling in a forest where nobody can hear it fall will actually make a sound. This article defines sound as mechanical vibrations traveling through the air or some other medium at a frequency to which the human ear is sensitive. Therefore, it would answer the above question in the affirmative, arguing that the mechanical vibrations composing a sound wave do exist whether or not anyone is present to hear them.

The science of sound is called acoustics, a word derived from the Greek *akoustos,* meaning "hearing." Beginning with its origins in the study of mechanical vibrations and the radiation of these vibrations through mechanical waves, acoustics has had important applications in almost every area of life. It has been fundamental to many developments in the arts—some of which, especially in the area of musical scales and instruments, took place after long experimentation by artists and were only much later explained as theory by scientists. For example, much of what is now known about architectural acoustics was actually learned by trial and error over centuries of experience and was only recently formalized into a science. Other applications of acoustic technology are in the study of geologic, atmospheric, and underwater phenomena. Psychoacoustics, the study of the physical effects of sound on biological systems, has been of interest since Pythagorus first heard the sounds of vibrating strings and of hammers hitting anvils in the 6th century BC, but the application of modern ultrasonic technology has only recently provided some of the most exciting developments in medicine. Even today, research continues into many aspects of the fundamental physical processes involved in waves and sound and into possible applications of these processes in modern life.

This article begins with an explanation of the physical properties of waves and sound and then discusses applications of acoustics to the areas of science, technology, the arts, and medicine. Also discussed are the related sciences of ultrasonics and infrasonics, which technically are not sound because they are not perceived by the ear but which have very important applications in modern technology.

Sound waves follow physical principles that can be applied to the study of all waves; these principles are discussed thoroughly in the *Macropædia* article MECHANICS. The article SENSORY RECEPTION explains in detail the physiological process of hearing—that is, receiving certain wave vibrations and interpreting them as sound.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 128 and 735, and the *Index.*

The article is divided into the following sections:

History of acoustics 556
  Early experimentation 556
  Measuring the speed of sound 557
  Modern advances 557
  Amplifying, recording, and reproducing 558
The propagation and perception of sound 558
  Plane waves 558
    Wavelength, period, and frequency
    Amplitude and intensity
    The speed of sound
  Circular and spherical waves 560
    Attenuation
    Diffraction
    Refraction
    Reflection
    Impedance
    Interference
    Moving sources and observers
  Standing waves 564
    In stretched strings
    In air columns
    In solid rods
    In nonharmonic systems
  Steady-state waves 567
    Spectral analysis
    Generation by musical instruments
    The human voice
    Noise
  Hearing 568
    Dynamic range of the ear
    The ear as spectrum analyzer
    Binaural perception
Applications of sound 570
  Electromechanical transducers 570
    Microphones
    Loudspeakers
  Sound recording 571
    The phonograph disc
    The audiotape
    The compact disc
  Architectural acoustics 572
    Reverberation time
    Acoustic criteria
    Acoustic problems
  Ultrasonics 573
    Transducers
    Applications in research
    Ranging and navigating
    The Doppler effect
    Materials testing
    High-intensity applications
    Chemical and electrical uses
    Medical applications
  Infrasonics 575
  Environmental noise 575
Bibliography 576

History of acoustics

EARLY EXPERIMENTATION

The origin of the science of acoustics is generally attributed to the Greek philosopher Pythagorus (6th century BC), whose experiments on the properties of vibrating strings that produce pleasing musical intervals were of such merit that they led to a tuning system that bears his name. Aristotle (4th century BC) correctly suggested that a sound wave propagates in air through motion of the air—a hypothesis based more on philosophy than on experimental physics; however, he also incorrectly suggested that high frequencies propagate faster than low frequencies—an error that persisted for many centuries. Vitruvius, a Roman architectural engineer of the 1st century BC, determined the correct mechanism for the transmission of sound waves, and he contributed substantially to the acoustic design of theatres. In the 6th century AD, the Roman philosopher Boethius documented several ideas relating science to music, including a suggestion that the human perception of pitch is related to the physical property of frequency.

The modern study of waves and acoustics is said to have originated with Galileo Galilei (1564–1642), who elevated to the level of science the study of vibrations and the correlation between pitch and frequency of the sound source. His interest in sound was inspired in part by his father, who was a mathematician, musician, and composer of some repute. Following Galileo's foundation work, progress in acoustics came relatively rapidly. The French mathematician Marin Mersenne studied the vibration of stretched

strings; the results of these studies were summarized in the three Mersenne's laws. Mersenne's *Harmonicorum Libri* (1636) provided the basis for modern musical acoustics. Later in the century Robert Hooke, an English physicist, first produced a sound wave of known frequency, using a rotating cog wheel as a measuring device. Further developed in the 19th century by the French physicist Félix Savart, and now commonly called Savart's disk, this device is often used today for demonstrations during physics lectures. In the late 17th and early 18th centuries, detailed studies of the relationship between frequency and pitch and of waves in stretched strings were carried out by the French physicist Joseph Sauveur, who provided a legacy of acoustic terms used to this day and first suggested the name acoustics for the study of sound.

The "bell-in-vacuum" experiment

One of the most interesting controversies in the history of acoustics involves the famous and often misinterpreted "bell-in-vacuum" experiment, which has become a staple of contemporary physics lecture demonstrations. In this experiment the air is pumped out of a jar in which a ringing bell is located; as air is pumped out, the sound of the bell diminishes until it becomes inaudible. As late as the 17th century many philosophers and scientists believed that sound propagated via invisible particles originating at the source of the sound and moving through space to affect the ear of the observer. The concept of sound as a wave directly challenged this view, but it was not established experimentally until the first bell-in-vacuum experiment was performed by Athanasius Kircher, a German scholar, who described it in his book *Musurgia Universalis* (1650). Even after pumping the air out of the jar, Kircher could still hear the bell, so he concluded incorrectly that air was not required to transmit sound. In fact, Kircher's jar was not entirely free of air, probably because of inadequacy in his vacuum pump. By 1660 the Anglo-Irish scientist Robert Boyle had improved vacuum technology to the point where he could observe sound intensity decreasing virtually to zero as air was pumped out. Boyle then came to the correct conclusion that a medium such as air is required for transmission of sound waves. Although this conclusion is correct, as an explanation for the results of the bell-in-vacuum experiment it is misleading. Even with the mechanical pumps of today, the amount of air remaining in a vacuum jar is more than sufficient to transmit a sound wave. The real reason for a decrease in sound level upon pumping air out of the jar is that the bell is unable to transmit the sound vibrations efficiently to the less dense air remaining, and that air is likewise unable to transmit the sound efficiently to the glass jar. Thus, the real problem is one of an impedance mismatch between the air and the denser solid materials—and not the lack of a medium such as air, as is generally presented in textbooks. Nevertheless, despite the confusion regarding this experiment, it did aid in establishing sound as a wave rather than as particles. (For further discussion of the concept of acoustic impedance, see below *The propagation and perception of sound.*)

MEASURING THE SPEED OF SOUND

Once it was recognized that sound is in fact a wave, measurement of the speed of sound became a serious goal. In the 17th century, the French scientist and philosopher Pierre Gassendi made the earliest known attempt at measuring the speed of sound in air. Assuming correctly that the speed of light is effectively infinite compared with the speed of sound, Gassendi measured the time difference between spotting the flash of a gun and hearing its report over a long distance on a still day. Although the value he obtained was too high—about 478.4 metres per second (1,569.6 feet per second)—he correctly concluded that the speed of sound is independent of frequency. In the 1650s, Italian physicists Giovanni Alfonso Borelli and Vincenzo Viviani obtained the much better value of 350 metres per second using the same technique. Their compatriot G.L. Bianconi demonstrated in 1740 that the speed of sound in air increases with temperature. The earliest precise experimental value for the speed of sound, obtained at the Academy of Sciences in Paris in 1738, was 332 metres per second—incredibly close to the presently accepted value, considering the rudimentary nature of the measuring tools of the day. A more recent value for the speed of sound, 331.45 metres per second (1,087.4 feet per second), was obtained in 1942; it was amended in 1986 to 331.29 metres per second at 0° C (1,086.9 feet per second at 32° F).

The speed of sound in air

The speed of sound in water was first measured by Daniel Colladon, a Swiss physicist, in 1826. Strangely enough, his primary interest was not in measuring the speed of sound in water but in calculating water's compressibility—a theoretical relationship between the speed of sound in a material and the material's compressibility having been established previously. Colladon came up with a speed of 1,435 metres per second at 8° C; the presently accepted value interpolated at that temperature is about 1,439 metres per second.

Two approaches were employed to determine the velocity of sound in solids. In 1808 Jean-Baptiste Biot, a French physicist, conducted direct measurements of the speed of sound in 1,000 metres of iron pipe by comparing it with the speed of sound in air. A better measurement had earlier been carried out by a German, Ernst Florenz Friedrich Chladni, using analysis of the nodal pattern in standing-wave vibrations in long rods.

MODERN ADVANCES

Simultaneous with these early studies in acoustics, theoreticians were developing the mathematical theory of waves required for the development of modern physics, including acoustics. In the early 18th century, the English mathematician Brook Taylor developed a mathematical theory of vibrating strings that agreed with previous experimental observations, but he was not able to deal with vibrating systems in general without the proper mathematical base. This was provided by Isaac Newton of England and Gottfried Wilhelm Leibniz of Germany, who, in pursuing other interests, independently developed the theory of calculus, which in turn allowed the derivation of the general wave equation by the French mathematician and scientist Jean Le Rond d'Alembert in the 1740s. The Swiss mathematicians Daniel Bernoulli and Leonhard Euler, as well as the Italian-French mathematician Joseph-Louis Lagrange, further applied the new equations of calculus to waves in strings and in the air. In the 19th century, Siméon-Denis Poisson of France extended these developments to stretched membranes, and the German mathematician Rudolf Friedrich Alfred Clebsch completed Poisson's earlier studies. A German experimental physicist, August Kundt, developed a number of important techniques for investigating properties of sound waves. These included the Kundt's tube, discussed below (see *The propagation and perception of sound: Standing waves*).

One of the most important developments in the 19th century involved the theory of vibrating plates. In addition to his work on the speed of sound in metals, Chladni had earlier introduced a technique of observing standing-wave patterns on vibrating plates by sprinkling sand onto the plates—a demonstration commonly used today. Perhaps the most significant step in the theoretical explanation of these vibrations was provided in 1816 by the French mathematician Sophie Germain, whose explanation was of such elegance and sophistication that errors in her treatment of the problem were not recognized until some 35 years later, by the German physicist Gustav Robert Kirchhoff.

Early spectral analysis

The analysis of a complex periodic wave into its spectral components was theoretically established early in the 19th century by Jean-Baptiste-Joseph Fourier of France and is now commonly referred to as the Fourier theorem. The German physicist Georg Simon Ohm first suggested that the ear is sensitive to these spectral components; his idea that the ear is sensitive to the amplitudes but not the phases of the harmonics of a complex tone is known as Ohm's law of hearing (distinguishing it from the more famous Ohm's law of electrical resistance).

Hermann von Helmholtz made substantial contributions to understanding the mechanisms of hearing and to the psychophysics of sound and music. His book *On the Sensations of Tone As a Physiological Basis for the Theory of Music* (1863) is one of the classics of acoustics. In addition, he constructed a set of resonators, covering much of

the audio spectrum, which were used in the spectral analysis of musical tones. The Prussian physicist Karl Rudolph Koenig, an extremely clever and creative experimenter, designed many of the instruments used for research in hearing and music, including a frequency standard and the manometric flame. The flame-tube device, used to render standing sound waves "visible," is still one of the most fascinating of physics classroom demonstrations. The English physical scientist John William Strutt, 3rd Baron Rayleigh, carried out an enormous variety of acoustic research; much of it was included in his two-volume treatise, *The Theory of Sound*, publication of which in 1877–78 is now thought to mark the beginning of modern acoustics. Much of Rayleigh's work is still directly quoted in contemporary physics textbooks.

The study of ultrasonics was initiated by the American scientist John LeConte, who in the 1850s developed a technique for observing the existence of ultrasonic waves with a gas flame. This technique was later used by the British physicist John Tyndall for the detailed study of the properties of sound waves. The piezoelectric effect, a primary means of producing and sensing ultrasonic waves, was discovered by the French physical chemist Pierre Curie and his brother Jacques in 1880. Applications of ultrasonics, however, were not possible until the development in the early 20th century of the electronic oscillator and amplifier, which were used to drive the piezoelectric element.

Among 20th-century innovators were the American physicist Wallace Sabine, considered to be the originator of modern architectural acoustics, and the Hungarian-born American physicist Georg von Békésy, who carried out experimentation on the ear and hearing and validated the commonly accepted place theory of hearing first suggested by Helmholtz. Békésy's book *Experiments in Hearing*, published in 1960, is the magnum opus of the modern theory of the ear.

AMPLIFYING, RECORDING, AND REPRODUCING

The earliest known attempt to amplify a sound wave was made by Athanasius Kircher, of "bell-in-vacuum" fame; Kircher designed a parabolic horn that could be used either as a hearing aid or as a voice amplifier. The amplification of body sounds became an important goal, and the first stethoscope was invented by a French physician, René Laënnec, in the early 19th century.

Attempts to record and reproduce sound waves originated with the invention in 1857 of a mechanical sound-recording device called the phonautograph by Édouard-Léon Scott de Martinville. The first device that could actually record and play back sounds was developed by the American inventor Thomas Alva Edison in 1877. Edison's phonograph employed grooves of varying depth in a cylindrical sheet of foil, but a spiral groove on a flat rotating disk was introduced a decade later by the German-born American inventor Emil Berliner in an invention he called the gramophone. Much significant progress in recording and reproduction techniques was made during the first half of the 20th century, with the development of high-quality electromechanical transducers and linear electronic circuits. The most important improvement on the standard phonograph record in the second half of the century was the compact disc, which employed digital techniques developed in mid-century that substantially reduced noise and increased the fidelity and durability of the recording.

The phonograph

The propagation and perception of sound

As stated above, sound is essentially a wave, so that a discussion of sound should begin with the properties of sound waves. There are two basic types of wave, transverse and longitudinal, differentiated by the way in which the wave is propagated. In a transverse wave, such as the wave generated in a stretched rope when one end is wiggled back and forth, the motion that constitutes the wave is perpendicular, or transverse, to the direction (along the rope) in which the wave is moving. An important family of transverse waves is generated by electromagnetic sources such as light or radio, in which the electric and magnetic fields constituting the wave oscillate perpendicular to the direction of propagation.

Sound propagates through air or other mediums as a longitudinal wave, in which the mechanical vibration constituting the wave occurs along the direction of propagation of the wave. A longitudinal wave can be created in a coiled spring by squeezing several of the turns together to form a compression and then releasing them, allowing the compression to travel the length of the spring. Air can be viewed as being composed of layers analogous to such coils, with a sound wave propagating as layers of air "push" and "pull" at one another much like the compression moving down the spring.

Sound as a longitudinal wave

A sound wave thus consists of alternating compressions and rarefactions, or regions of high pressure and low pressure, moving at a certain speed. Put another way, it consists of a periodic (that is, oscillating or vibrating) variation of pressure occurring around the equilibrium pressure prevailing at a particular time and place. Equilibrium pressure and the sinusoidal variations caused by passage of a pure sound wave (that is, a wave of a single frequency) are represented in Figure 1A and 1B, respectively.

PLANE WAVES

A discussion of sound waves and their propagation can begin with an examination of a plane wave of a single frequency passing through the air. A plane wave is a wave that propagates through space as a plane, rather than as a sphere of increasing radius. As such, it is not perfectly representative of sound (see below *Circular and spherical waves*). A wave of single frequency would be heard as a pure sound such as that generated by a tuning fork that has been lightly struck. As a theoretical model, it helps to elucidate many of the properties of a sound wave.

Wavelength, period, and frequency. Figure 1C is another representation of the sound wave illustrated in Figure 1B. As represented by the sinusoidal curve, the pressure variation in a sound wave repeats itself in space over a specific distance. This distance is known as the wavelength of the sound, usually measured in metres and represented by λ. As the wave propagates through the air, one full wavelength takes a certain time period to pass a specific point in space; this period, represented by T, is usually measured in fractions of a second. In addition, during each one-second time interval, a certain number of wavelengths pass a point in space. Known as the frequency of the sound wave, the number of wavelengths passing per second is traditionally measured in hertz or kilohertz and is represented by f.

There is an inverse relation between a wave's frequency and its period, such that

$$fT = 1 \text{ or } f = \frac{1}{T}. \qquad (1)$$

This means that sound waves with high frequencies have short periods, while those with low frequencies have long periods. For example, a sound wave with a frequency of 20 hertz would have a period of 0.05 second (*i.e.*, 20 wavelengths/second $\times$ 0.05 second/wavelength = 1), while a sound wave of 20 kilohertz would have a period of 0.00005 second (20,000 wavelengths/second $\times$ 0.00005 second/wavelength = 1). Between 20 hertz and 20 kilohertz lies the frequency range of hearing for humans. The physical property of frequency is perceived physiologically as pitch, so that the higher the frequency, the higher the perceived pitch. There is also a relation between the wavelength of a sound wave, its frequency or period, and the speed of the wave (S), such that

The frequency range of human hearing

$$S = f\lambda = \frac{\lambda}{T}. \qquad (2)$$

Amplitude and intensity. *Mathematical values.* The equilibrium value of pressure, represented by the evenly spaced lines in Figure 1A and by the axis of the graph in Figure 1C, is equal to the atmospheric pressure that would prevail in the absence of the sound wave. With passage of the compressions and rarefactions that constitute the sound wave, there would occur a fluctuation above and below atmospheric pressure. The magnitude of this fluctu-

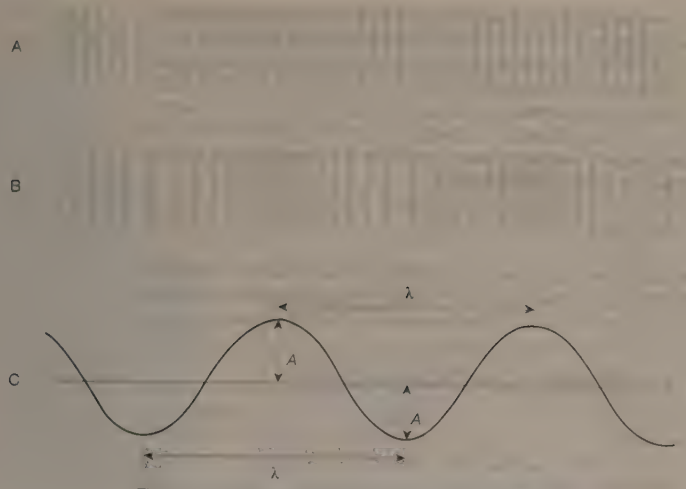


Figure 1: *Graphic representations of a sound wave.*
(A) Air at equilibrium, in the absence of a sound wave; (B) compressions and rarefactions that constitute a sound wave; (C) transverse representation of the wave, showing amplitude (A) and wavelength (λ).

ation from equilibrium is known as the amplitude of the sound wave; measured in pascals, or newtons per square metre, it is represented by the letter A. The displacement or disturbance of a plane sound wave can be described mathematically by the general equation for wave motion, which is written in simplified form as:

$$y(x,t) = A \sin 2\pi(ft - x/\lambda). \quad (3)$$

This means that the size of the disturbance (y), for any value of time (t) or distance (x) away from the source of the sound, is equal to the amplitude (A) times the sine of 2π, times a quantity equal to the frequency (f) times elapsed time (t) minus the distance from origin (x) divided by the wavelength (λ).

The amplitude of a sound wave determines its intensity, which in turn is perceived by the ear as loudness. Acoustic intensity is defined as the average rate of energy transmission per unit area perpendicular to the direction of propagation of the wave. Its relation with amplitude can be written as

$$I = \frac{A^2}{2\rho S}, \quad (4)$$

where ρ is the equilibrium density of the air (measured in kilograms per cubic metre) and S is the speed of sound (in metres per second). Intensity (I) is measured in watts per square metre, the watt being the standard unit of power in electrical or mechanical usage.

The value of atmospheric pressure under "standard atmospheric conditions" is generally given as about 10^5 pascals, or 10^5 newtons per square metre. The minimum amplitude of pressure variation that can be sensed by the human ear is about 10^{-5} pascal, and the pressure amplitude at the threshold of pain is about 10 pascals, so the pressure variation in sound waves is very small compared with the pressure of the atmosphere. Under these conditions a sound wave propagates in a linear manner—that is, it continues to propagate through the air with very little loss, dispersion, or change of shape. However, when the amplitude of the wave reaches about 100 pascals (approximately one one-thousandth the pressure of the atmosphere), significant nonlinearities develop in the propagation of the wave.

Nonlinearity of high-amplitude waves

Nonlinearity arises from the peculiar effects on air pressure caused by a sinusoidal displacement of air molecules. When the vibratory motion constituting a wave is small, the increase and decrease in pressure are also small and are very nearly equal. But when the motion of the wave is large, each compression generates an excess pressure of greater amplitude than the decrease in pressure caused by each rarefaction. This can be predicted by the ideal gas law, which states that increasing the volume of a gas by one-half decreases its pressure by only one-third, while decreasing its volume by one-half increases the pressure by a factor of two. The result is a net excess in pressure—a phenomenon that is significant only for waves with amplitudes above about 100 pascals.

The decibel scale. The ear mechanism is able to respond to both very small and very large pressure waves by virtue of being nonlinear; that is, it responds much more efficiently to sounds of very small amplitude than to sounds of very large amplitude. Because of the enormous nonlinearity of the ear in sensing pressure waves, a nonlinear scale is convenient in describing the intensity of sound waves. Such a scale is provided by the sound intensity level, or decibel level, of a sound wave, which is defined by the equation

$$L = 10 \log \left(\frac{I}{I_0}\right). \quad (5)$$

Here L represents decibels, which correspond to an arbitrary sound wave of intensity I, measured in watts per square metre. The reference intensity I_0, corresponding to a level of 0 decibels, is approximately the intensity of a wave of 1,000 hertz frequency at the threshold of hearing—about 10^{-12} watt per square metre. Because the decibel scale mirrors the function of the ear more accurately than a linear scale, it has several advantages in practical use; these are discussed in *Hearing*, below.

A fundamental feature of this type of logarithmic scale is that each unit of increase in the decibel scale corresponds to an increase in absolute intensity by a constant multiplicative factor. Thus, an increase in absolute intensity from 10^{-12} to 10^{-11} watt per square metre corresponds to an increase of 10 decibels, as does an increase from 10^{-1} to 1 watt per square metre. The correlation between the absolute intensity of a sound wave and its decibel level is shown in Table 1, along with examples of sounds at each level. When the defining level of 0 decibel (10^{-12} watt per square metre) is taken to be at the threshold of hearing for a sound wave with a frequency of 1,000 hertz, then 130 decibels (10 watts per square metre) corresponds to the threshold of feeling, or the threshold of pain. (Sometimes the threshold of pain is given as 120 decibels, or 1 watt per square metre.)

Table 1: Sound Levels for Nonlinear (Decibel) and Linear (Intensity) Scales

decibels	intensity*	type of sound
130	10	artillery fire at close proximity (threshold of pain)
120	1	amplified rock music; near jet engine
110	10^{-1}	loud orchestral music, in audience
100	10^{-2}	electric saw
90	10^{-3}	bus or truck interior
80	10^{-4}	automobile interior
70	10^{-5}	average street noise; loud telephone bell
60	10^{-6}	normal conversation; business office
50	10^{-7}	restaurant; private office
40	10^{-8}	quiet room in home
30	10^{-9}	quiet lecture hall; bedroom
20	10^{-10}	radio, television, or recording studio
10	10^{-11}	soundproof room
0	10^{-12}	absolute silence (threshold of hearing)

*In watts per square metre.

Although the decibel scale is nonlinear, it is directly measurable, and sound-level meters are available for that purpose. Sound levels for audio systems, architectural acoustics, and other industrial applications are most often quoted in decibels.

The speed of sound. *In gases.* For longitudinal waves such as sound, wave velocity is in general given as the square root of the ratio of the elastic modulus of the medium (that is, the ability of the medium to be compressed by an external force) to its density:

$$S = \sqrt{\frac{B}{\rho}}. \quad (6)$$

Here ρ is the density and B the bulk modulus (the ratio of the applied pressure to the change in volume per unit volume of the medium). In gas mediums this equation is modified to

$$S = \sqrt{\frac{1}{\rho K}}. \quad (7)$$

where K is the compressibility of the gas. Compressibility (K) is the reciprocal of the bulk modulus (B), as in

$$K = \frac{1}{B}. \quad (8)$$

Using the appropriate gas laws, wave velocity can be calculated in two ways, in relation to pressure or in relation to temperature:

$$S = \sqrt{\frac{\gamma p}{\rho}}; \quad (9)$$

or

$$S = \sqrt{\frac{\gamma R\theta}{M}}. \quad (10)$$

Here p is the equilibrium pressure of the gas in pascals, ρ is its equilibrium density in kilograms per cubic metre at pressure p, θ is absolute temperature in kelvins, R is the gas constant per mole, M is the molecular weight of the gas, and γ is the ratio of the specific heat at a constant pressure to the specific heat at a constant volume,

$$\gamma = \sqrt{\frac{C_p}{C_v}}. \quad (11)$$

Values for γ for various gases are given in many physics textbooks and reference works. The speed of sound in several different gases, including air, is given in Table 2.

Table 2: Speed of Sound in Selected Gases

gas	speed	
	(metres/second)	(feet/second)
Helium, at 0° C (32° F)	965	3,165
Nitrogen, at 0° C	334	1,096
Oxygen, at 0° C	316	1,036
Carbon dioxide, at 0° C	259	850
Air, dry, at 0° C	345.45	1,133
Steam, at 134° C (273° F)	494	1,620

Velocity independent of pressure and frequency

Equation (10) states that the speed of sound depends only on absolute temperature and not on pressure, since, if the gas behaves as an ideal gas, then its pressure and density, as shown in equation (9), will be proportional. This means that the speed of sound does not change between locations at sea level and high in the mountains and that the pitch of wind instruments at the same temperature is the same anywhere. In addition, both equations (9) and (10) are independent of frequency, indicating that the speed of sound is in fact the same at all frequencies—that is, there is no dispersion of a sound wave as it propagates through air. One assumption here is that the gas behaves as an ideal gas. However, gases at very high pressures no longer behave like an ideal gas, and this results in some absorption and dispersion. In such cases equations (9) and (10) must be modified, as they are in advanced books on the subject.

In liquids. For a liquid medium, the appropriate modulus is the bulk modulus, so that the speed of sound is equal to the square root of the ratio of the bulk modulus (B) to the equilibrium density (ρ), as shown in equation (6) above. The speed of sound in liquids under various conditions is given in Table 3. The speed of sound in liquids varies slightly with temperature—a variation that is accounted for by empirical corrections to equation (6), as is indicated in the values given for water in Table 3.

Table 3: Speed of Sound in Selected Liquids
(at one atmosphere pressure)

liquid	speed	
	(metres/second)	(feet/second)
Pure water, at 0° C (32° F)	1,402.3	4,600
Pure water, at 30° C (86° F)	1,509.0	4,950
Pure water, at 50° C (122° F)	1,542.5	5,060
Pure water, at 70° C (158° F)	1,554.7	5,100
Pure water, at 100° C (212° F)	1,543.0	5,061
Salt water, at 0° C	1,449.4	4,754
Salt water, at 30° C	1,546.2	5,072
Methyl alcohol, at 20° C (68° F)	1,121.2	3,678
Mercury, at 20° C	1,451.0	4,760

In solids. For a long, thin solid the appropriate modulus is the Young's, or stretching, modulus (the ratio of the applied stretching force per unit area of the solid to the resulting change in length per unit length; named for the English physicist and physician Thomas Young). The speed of sound, therefore, is

$$S = \sqrt{\frac{Y}{\rho}}, \quad (12)$$

where Y is the Young's modulus and ρ is the density. Table 4 gives the speed of sound in representative solids.

In the case of a three-dimensional solid, in which the wave is traveling outward in spherical waves, the above expression becomes more complicated. Both the shear modulus, represented by η, and the bulk modulus B play a role in the elasticity of the medium:

$$S = \sqrt{\frac{(B + 4\eta/3)}{\rho}}. \quad (13)$$

Table 4: Speed of Sound in Selected Solids

solid	speed	
	(metres/second)	(feet/second)
Aluminum, rolled	5,000	16,500
Copper, rolled	3,750	12,375
Iron, cast	4,480	14,784
Lead	1,210	3,993
Pyrex (trademark)	5,170	17,061
Lucite (trademark)	1,840	6,072

CIRCULAR AND SPHERICAL WAVES

The above discussion of the propagation of sound waves begins with a simplifying assumption that the wave exists as a plane wave. In most real cases, however, a wave originating at some source does not move in a straight line but expands in a series of spherical wavefronts. The fundamental mechanism for this propagation is known as

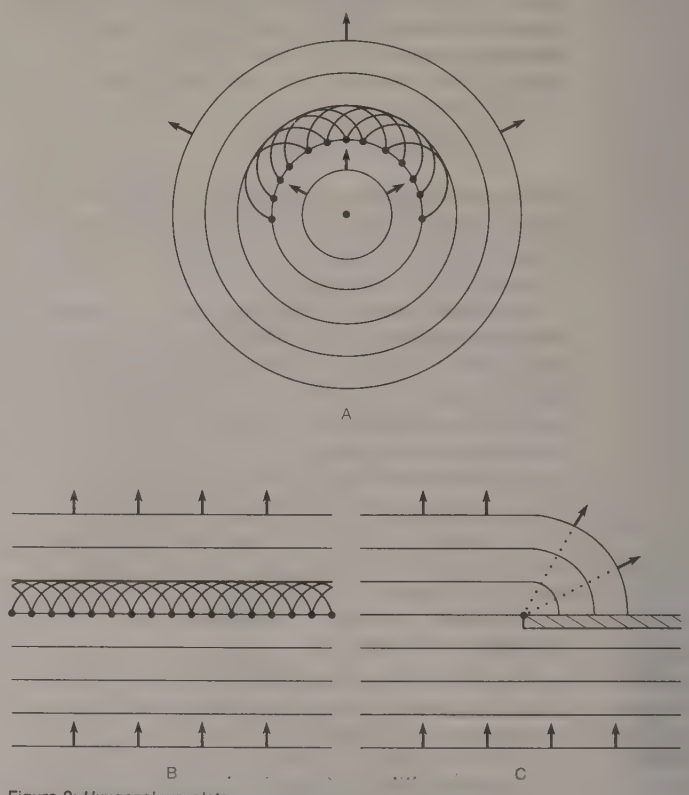


Figure 2: *Huygens' wavelets.*
Originating along the fronts of (A) circular waves and (B) plane waves, wavelets recombine to produce the propagating wave front. (C) The diffraction of sound around a corner arising from Huygens' wavelets.

Huygens' wavelets

Huygens' principle, according to which every point on a wave is a source of spherical waves in its own right. The result is a Huygens' wavelet construction, illustrated in Figure 2A and 2B for a two-dimensional plane wave and circular wave. The insightful point suggested by the Dutch physicist Christiaan Huygens is that all the wavelets of Figure 2A and 2B, including those not shown but originating between those that are shown, form a new coherent wave that moves along at the speed of sound to form the next wave in the sequence. In addition, just as the wavelets add up in the forward direction to create a new wavefront, they also cancel one another, or interfere destructively, in the backward direction, so that the waves continue to propagate only in the forward direction.

The principle behind the adding up of Huygens' wavelets, involving a fundamental difference between matter and waves, is known as the principle of superposition. The old saying that no two things can occupy the same space at the same time is correct when applied to matter, but it does not apply to waves. Indeed, an infinite number of waves can occupy the same space at the same time; furthermore, they do this without affecting one another, so that each wave retains its own character independent of how many other waves are present at the same point and time. A radio or television antenna can receive the signal of any single frequency to which it is tuned, unaffected by the existence of any others. Likewise, the sound waves of two people talking may cross each other, but the sound of each voice is unaffected by the waves' having been simultaneously at the same point.

Superposition plays a key role in many of the wave properties of sound discussed in this section. It is also fundamental to the addition of Fourier components of a wave in order to obtain a complex wave shape (see below *Steady-state waves*).

Attenuation. *The inverse square law.* A plane wave of a single frequency in theory will propagate forever with no change or loss. This is not the case with a circular or spherical wave, however. One of the most important properties of this type of wave is a decrease in intensity as the wave propagates. The mathematical explanation of this principle, which derives as much from geometry as from physics, is known as the inverse square law.

Decreasing intensity of an expanding wave

As a circular wave front (such as that created by dropping a stone onto a water surface) expands, its energy is distributed over an increasingly larger circumference. The intensity, or energy per unit of length along the circumference of the circle, will therefore decrease in an inverse relationship with the growing radius of the circle, or distance from the source of the wave. In the same way, as a spherical wave front expands, its energy is distributed over a larger and larger surface area. Because the surface area of a sphere is proportional to the square of its radius, the intensity of the wave is inversely proportional to the square of the radius. This geometric relation between the growing radius of a wave and its decreasing intensity is what gives rise to the inverse square law.

The decrease in intensity of a spherical wave as it propagates outward can also be expressed in decibels. Each factor of two in distance from the source leads to a decrease in intensity by a factor of four. For example, a factor of four decrease in a wave's intensity is equivalent to a decrease of six decibels, so that a spherical wave attenuates at a rate of six decibels for each factor of two increase in distance from the source. If a wave is propagating as a hemispherical wave above an absorbing surface, the intensity will be further reduced by a factor of two near the surface because of the lack of contributions of Huygens' wavelets from the missing hemisphere. Thus, the intensity of a wave propagating along a level, perfectly absorbent floor falls off at the rate of 12 decibels for each factor of two in distance from the source. This additional attenuation leads to the necessity of sloping the seats of an auditorium in order to retain a good sound level in the rear.

Sound absorption. In addition to the geometric decrease in intensity caused by the inverse square law, a small part of a sound wave is lost to the air or other medium through various physical processes. One important process is the direct conduction of the vibration into the medium as heat, caused by the conversion of the coherent molecular motion of the sound wave into incoherent molecular motion in the air or other absorptive material. Another cause is the viscosity of a fluid medium (*i.e.*, a gas or liquid). These two physical causes combine to produce the classical attenuation of a sound wave. This type of attenuation is proportional to the square of the sound wave's frequency, as expressed in the formula α/f^2, where α is the attenuation coefficient of the medium and f is the wave frequency. The amplitude of an attenuated wave is then given by

$$A(x) = A_o\, e^{-\alpha x}, \tag{14}$$

where A_o is the original amplitude of the wave and $A(x)$ is the amplitude after it has propagated a distance x through the medium.

Table 5 gives sound-absorption coefficients for several gases. The magnitudes of the coefficients indicate that, although attenuation is rather small for audible frequencies, it can become extremely large for high-frequency ultrasonic waves. Attenuation of sound in air also varies with temperature and humidity.

Table 5: Attenuation of Sound in Selected Fluids

fluid	attenuation coefficient $10^{13}\alpha/f^2$ (s^2/cm)
Helium	52.5
Hydrogen	16.9
Nitrogen	133.0
Oxygen	165.0
Air	137.0
Carbon dioxide	140.0
Water, at 0° C (32° F)	0.569
Water, at 20° C (68° F)	0.253
Water, at 80° C (176° F)	0.079
Mercury, at 25° C (77° F)	0.057
Methyl alcohol, at 30° C (86° F)	0.302

Greater propagation of sound in solids and liquids

Because less sound is absorbed in solids and liquids than in gases, sounds can propagate over much greater distances in these mediums. For instance, the great range over which certain sea mammals can communicate is made possible partially by the low attenuation of sound in water. In addition, because absorption increases with frequency, it becomes very difficult for ultrasonic waves to penetrate a dense medium. This is a persistent limitation on the development of high-frequency ultrasonic applications.

Most sound-absorbing materials are nonlinear, in that they do not absorb the same fraction of acoustical waves of all frequencies. In architectural acoustics, an enormous effort is expended to use construction materials that absorb undesirable frequencies but reflect desired frequencies. Absorption of undesirable sound, such as that from machines in factories, is critical to the health of workers, and noise control in architectural and industrial acoustics has expanded to become an important field of environmental engineering.

Diffraction. A direct result of Huygens' wavelets is the property of diffraction, the capacity of sound waves to bend around corners and to spread out after passing through a small hole or slit. If a barrier is placed in the path of half of a plane wave, as shown in Figure 2C, the part of the wave passing just by the barrier will propagate in a series of Huygens' wavelets, causing the wave to spread into the shadow region behind the barrier. In light waves, wavelengths are very small compared with the size of everyday objects, so that very little diffraction occurs and a relatively clear shadow can be formed. The wavelengths of sound waves, on the other hand, are more nearly equal to the size of everyday objects, so that they readily diffract.

Diffraction of sound is helpful in the case of audio systems, in which sound emanating from loudspeakers spreads out and reflects off of walls to fill a room. It is also the reason why "sound beams" cannot generally be produced like light beams. On the other hand, the ability of a sound wave to diffract decreases as frequency rises and wavelength shrinks. This means that the lower frequencies of a voice bend around a corner more readily than the higher frequencies, giving the diffracted voice a "muffled"

sound. Also, because the wavelengths of ultrasonic waves become extremely small at high frequencies, it is possible to create a beam of ultrasound. Ultrasonic beams have become very useful in modern medicine (see below *Applications of sound*).

Scattering of sound

The scattering of a sound wave is a reflection of some part of the wave off of an obstacle around which the rest of the wave propagates and diffracts. The way in which the scattering occurs depends upon the relative size of the obstacle and the wavelength of the scattering wave. If the wavelength is large in relation to the obstacle, then the wave will pass by the obstacle virtually unaffected. In this case, the only part of the wave to be scattered will be the tiny part that strikes the obstacle; the rest of the wave, owing to its large wavelength, will diffract around the obstacle in a series of Huygens' wavelets and remain unaffected. If the wavelength is small in relation to the obstacle, the wave will not diffract strongly, and a shadow will be formed similar to the optical shadow produced by a small light source. In extreme cases, arising primarily with high-frequency ultrasound, the formalism of ray optics often used in lenses and mirrors can be conveniently employed.

If the size of the obstacle is the same order of magnitude as the wavelength, diffraction may occur, and this may result in interference among the diffracted waves. This would create regions of greater and lesser sound intensity, called acoustic shadows, after the wave has propagated past the obstacle. Control of such acoustic shadows becomes important in the acoustics of auditoriums.

Refraction. Diffraction involves the bending or spreading out of a sound wave in a single medium, in which the speed of sound is constant. Another important case in which sound waves bend or spread out is called refraction. This phenomenon involves the bending of a sound wave owing to changes in the wave's speed. Refraction is the reason why ocean waves approach a shore parallel to the beach and why glass lenses can be used to focus light waves. An important refraction of sound is caused by the natural temperature gradient of the atmosphere. Under normal conditions the Sun heats the Earth and the Earth heats the adjacent air. The heated air then cools as it rises, creating a gradient in which atmospheric temperature decreases with elevation by an amount known as the adiabatic lapse rate. Because sound waves propagate faster in warm air, they travel faster closer to the Earth. This greater speed of sound in warmed air near the ground creates Huygens' wavelets that also spread faster near the ground. Because a sound wave propagates in a direction perpendicular to the wave front formed by all the Huygens' wavelets, sound under these conditions tends to refract upward and become "lost." The sound of thunder created by lightning may be refracted upward so strongly that a shadow region is created in which the lightning can be seen but the thunder cannot be heard. This typically occurs at a horizontal distance of about 22.5 kilometres (14 miles) from a lightning bolt about 4 kilometres high.

Greater propagation of sound at night

At night or during periods of dense cloud cover, a temperature inversion occurs; the temperature of the air increases with elevation, and sound waves are refracted back down to the ground. Temperature inversion is the reason why sounds can be heard much more clearly over longer distances at night than during the day—an effect often incorrectly attributed to the psychological result of nighttime quiet. The effect is enhanced if the sound is propagated over water, allowing sound to be heard remarkably clearly over great distances.

Refraction is also observable on windy days. Wind, moving faster at greater heights, causes a change in the effective speed of sound with distance above ground. When one speaks with the wind, the sound wave is refracted back down to the ground, and one's voice is able to "carry" farther than on a still day. When one speaks into the wind, however, the sound wave is refracted upward, away from the ground, and the voice is "lost."

Another example of sound refraction occurs in the ocean. Under normal circumstances the temperature of the ocean decreases with depth, resulting in the downward refraction of a sound wave originating under water—just the opposite of the shadow effect in air described above. Many marine biologists believe that this refraction enhances the propagation of the sounds of marine mammals such as dolphins and whales, allowing them to communicate with one another over enormous distances. For ships such as submarines located near the surface of the water, this refraction creates shadow regions, limiting their ability to locate distant vessels.

Reflection. A property of waves and sound quite familiar in the phenomenon of echoes is reflection. This plays a critical role in room and auditorium acoustics, in large part determining the adequacy of a concert hall for musical performance or other functions. In the case of light waves passing from air through a glass plate, close inspection shows that some of the light is reflected at each of the air-glass interfaces while the rest passes through the glass. This same phenomenon occurs whenever a sound wave passes from one medium into another—that is, whenever the speed of sound changes or the way in which the sound propagates is substantially modified.

The direction of propagation of a wave is perpendicular to the front formed by all the Huygens' wavelets. As a plane wave reflects off some reflector, the reflector directs the wave fronts formed by the Huygens' wavelets just as a light reflector directs light "rays." The same law of reflection is followed for both sound and light, so that focusing a sound wave is equivalent to focusing a light ray.

Focusing of sound

Reflectors of appropriate shape are used for a variety of purposes or effects. For example, a parabolic reflector will focus a parallel wave of sound onto a specific point, allowing a very weak sound to be more easily heard. Such reflectors are used in parabolic microphones to collect sound from a distant source or to choose a location from which sound is to be observed and then focus it onto a microphone. An elliptical shape, on the other hand, can be used to focus sound from one point onto another—an arrangement called a whispering chamber. Domes in cathedrals and capitols closely approximate the shape of an ellipse, so that such buildings often possess focal points and function as a type of whispering chamber. Concert halls must avoid the smooth, curved shape of ellipses and parabolas, because strong echoes or focusing of sound from one point to another are undesirable in an auditorium.

Impedance. One of the important physical characteristics relating to the propagation of sound is the acoustic impedance of the medium in which the sound wave travels. Acoustic impedance (Z) is given by the ratio of the wave's acoustic pressure (p) to its volume velocity (U):

$$Z = \frac{p}{U} \qquad (15)$$

Like its analogue, electrical impedance (or electrical resistance), acoustic impedance is a measure of the ease with which a sound wave propagates through a particular medium. Also like electrical impedance, acoustic impedance involves several different effects applying to different situations. For example, specific acoustic impedance (z), the ratio of acoustic pressure to particle speed, is an inherent property of the medium and of the nature of the wave. Acoustic impedance, the ratio of pressure to volume velocity, is equal to the specific acoustic impedance per unit area. Specific acoustic impedance is useful in discussing waves in confined mediums, such as tubes and horns. For the simplest case of a plane wave, specific acoustic impedance is the product of the equilibrium density (ρ) of the medium and the wave speed (S):

$$z = \rho S. \qquad (16)$$

The unit of specific acoustic impedance is the pascal second per metre, often called the rayl, after Lord Rayleigh. The unit of acoustic impedance is the pascal second per cubic metre, called an acoustic ohm, by analogy to electrical impedance.

Impedance mismatch. Mediums in which the speed of sound is different generally have differing acoustic impedances, so that, when a sound wave strikes an interface between the two, it encounters an impedance mismatch. As a result, some of the wave reflects while some is transmitted into the second medium. This situation is similar to that of a light wave entering glass, where a close

inspection will show that in general some light is reflected while most of the light is transmitted into the glass. In the case of the bell-in-vacuum experiment described above (see *History of acoustics*), the impedance mismatches between the bell and the air and between the air and the jar result in very little transmission of sound when the air is at low pressure.

Reducing the impedance mismatch

The efficiency with which a sound source radiates sound is enhanced by reducing the impedance mismatch between the source and the outside air. For example, if a tuning fork is struck and held in the air, it will be nearly inaudible because of the inability of the vibrations of the tuning fork to radiate efficiently to the air. Touching the tuning fork to a wooden plate such as a tabletop will enhance the sound by providing better coupling between the vibrating tuning fork and the air. This principle is used in the violin and the piano, in which the vibrations of the strings are transferred first to the back and belly of the violin or to the piano's sounding board, and then to the air.

Acoustic filtration. Filtration of sound plays an important part in the design of air-handling systems. In order to attenuate the level of sound from blower motors and other sources of vibration, regions of larger or smaller cross-sectional area are inserted into air ducts, as illustrated in Figure 3. The impedance mismatch introduced into a duct by a change in the area of the duct or by the addition of a side branch reflects undesirable frequencies, as determined by the size and shape of the variation. A region of either larger or smaller area will function as a low-pass filter, reflecting high frequencies; an opening or series of openings will function as a high-pass filter, removing low frequencies. Some automobile mufflers make use of this type of filter.

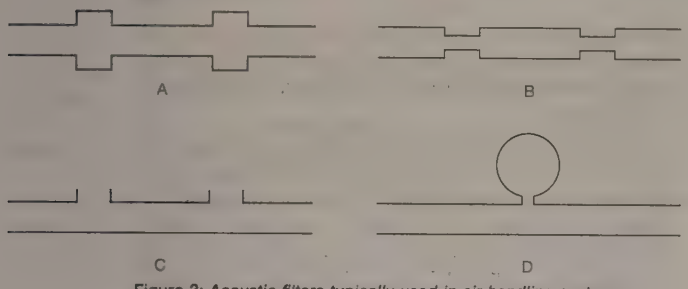


Figure 3: *Acoustic filters typically used in air-handling systems.* (A) and (B) Low-pass filters; (C) a high-pass filter; (D) a band-pass filter, which actually filters out vibrations within a narrow frequency range (see text).

A connected spherical cavity, forming what is called a band-pass filter, actually functions as a type of band absorber or notch filter, removing a band of frequencies around the resonant frequency of the cavity (see below, *Standing waves: The Helmholtz resonator*).

Interference. *Constructive and destructive.* The particular manner in which sound waves can combine is known as interference. Two identical waves in the same place at the same time can interfere constructively if they are in phase or destructively if they are out of phase. "Phase" is a term that refers to the time relationship between two periodic signals. "In phase" means that they are vibrating together, while "out of phase" means that their vibrations are opposite. Opposite vibrations added together cancel each other.

Phase

Constructive interference leads to an increase in the amplitude of the sum wave, while destructive interference can lead to the total cancellation of the contributing waves. An interesting example of both interference and diffraction of sound, called the "speaker and baffle" experiment, involves a small loudspeaker and a large, square wooden sheet with a circular hole in it the size of the speaker. When music is played on the loudspeaker, sound waves from the front and back of the speaker, which are out of phase, diffract into the entire region around the speaker. The two waves interfere destructively and cancel each other, particularly at very low frequencies, where the wavelength is longest and the diffraction is thus greatest. When the speaker is held up behind the baffle, though, the sounds can no longer diffract and mix while they are out of phase, and as a consequence the intensity increases enormously. This experiment illustrates why loudspeakers are often mounted in boxes, so that the sound from the back cannot interfere with the sound from the front. In a home stereo system, when two speakers are wired properly, their sound waves are in phase along an antinodal line between the two speakers and in the area of best listening. If the two speakers are wired incorrectly—the wires being reversed on one of the speakers—their waves will be out of phase in the area of best listening and will interfere destructively—especially at low frequencies, so that the bass frequencies will be strongly attenuated.

One possible application of destructive interference is in industrial noise control. This would involve sensing the ambient sound in a workplace, electronically reproducing a sound with the opposite phase, and then introducing that sound into the environment so that it would interfere destructively with the ambient sound and reduce the overall sound level.

Beats. An important occurrence of the interference of waves is in the phenomenon of beats. In the simplest case, beats result when two sinusoidal sound waves of equal amplitude and very nearly equal frequencies mix. The frequency of the resulting sound (F) would be the average of the two original frequencies (f_1 and f_2):

$$F = \frac{f_1 + f_2}{2}. \tag{17}$$

The amplitude or intensity of the combined signal would rise and fall at a rate (f_b) equal to the difference between the two original frequencies,

$$f_b = f_1 - f_2, \tag{18}$$

where f_1 is greater than f_2.

Beats are useful in tuning musical instruments to each other: the farther the instruments are out of tune, the faster the beats. Other types of beats are also of interest. Second-order beats occur between the two notes of a mistuned octave, and binaural beats involve beating between tones presented separately to the two ears, so that they do not mix physically.

Moving sources and observers. *The Doppler effect.* The Doppler effect is a change in the frequency of a tone that occurs by virtue of relative motion between the source of sound and the observer. When the source and the observer are moving closer together, the perceived frequency is higher than the normal frequency, or the frequency heard when the observer is at rest with respect to the source. When the source and the observer are moving farther apart, the perceived frequency is lower than the normal frequency. For the case of a moving source, one example is the falling frequency of a train whistle as the train passes a crossing. In the case of a moving observer, a passenger on the train would hear the warning bells at the crossing drop in frequency as the train sped by.

Perceived frequency

For the case of motion along a line, where the source moves with speed v_s and the observer moves with speed v_o through still air in which the speed of sound is S, the general equation describing the change in frequency heard by the observer is

$$f_o = f_s \frac{S + v_o}{S - v_s}. \tag{19}$$

In this equation the speeds of the source and the observer will be negative if the relative motion between the source and observer is moving them apart, and they will be positive if the source and observer are moving together.

From this equation, it can be deduced that a Doppler effect will always be heard as long as the relative speed between the source and observer is less than the speed of sound. The speed of sound is constant with respect to the air in which it is propagating, so that, if the observer moves away from the source at a speed greater than the speed of sound, nothing will be heard. If the source and the observer are moving with the same speed in the same direction, v_o and v_s will be equal in magnitude but with the opposite sign; the frequency of the sound will therefore remain unchanged, like the sound of a train whistle as heard by a passenger on the moving train.

Shock waves. If the speed of the source is greater than the speed of sound, another type of wave phenomenon will occur: the sonic boom. A sonic boom is a type of shock wave that occurs when waves generated by a source over a period of time add together coherently, creating an unusually strong sum wave. An analogue to a sonic boom is the V-shaped bow wave created in water by a motorboat when its speed is greater than the speed of the waves. In the case of an aircraft flying faster than the speed of sound (about 1,230 kilometres per hour, or 764 miles per hour), the shock wave takes the form of a cone in three-dimensional space called the Mach cone. The Mach number is defined as the ratio of the speed of the aircraft to the speed of sound. The higher the Mach number—that is, the faster the aircraft—the smaller the angle of the Mach cone.

STANDING WAVES

This section focuses on waves in bounded mediums—in particular, standing waves in such systems as stretched strings, air columns, and stretched membranes. The principles discussed here are directly applicable to the operation of string and wind instruments.

Combining of component waves

When two identical waves move in opposite directions along a line, they form a standing wave—that is, a wave form that does not travel through space or along a string even though (or because) it is made up of two oppositely traveling waves. The resulting standing wave is sinusoidal, like its two component waves, and it oscillates at the same frequency. An easily visualized standing wave can be created by stretching a rubber band between two fixed points, displacing its centre slightly, and releasing it so that it vibrates back and forth between two extremes. In musical instruments, a standing wave can be generated by driving the oscillating medium (such as the reeds of a woodwind) at one end; the standing waves are then created not by two separate component waves but by the original wave and its reflections off the ends of the vibrating system.

In stretched strings. *Fundamentals and harmonics.* For a stretched string of a given mass per unit length (μ) and under a given tension (F), the speed (v) of a wave in the string is given by the following equation:

$$v = \sqrt{\frac{F}{\mu}}. \tag{20}$$

When a string of a given length (L) is plucked gently in the middle, a vibration is produced with a wavelength (λ) that is twice the length of the string:

$$\lambda = 2L. \tag{21}$$

The frequency (f_1) of this vibration can then be obtained by the following adaptation of equation (2):

$$f_1 = \frac{v}{\lambda} = \left(\frac{1}{2L}\right)\sqrt{\frac{F}{\mu}}. \tag{22}$$

As the vibration that has the lowest frequency for that particular type and length of string under a specific tension, this frequency is known as the fundamental, or first harmonic.

Additional standing waves can be created in a stretched string; the three simplest are represented graphically in Figure 4. At the top is a representation of the fundamental, which is labeled $n = 1$. Because a string must be stretched by holding it in place at its ends, each end is fixed, and there can be no motion of the string at these points. The ends are called nodal points, or nodes, and labeled N. The shape of the string at the extreme positions in its oscillation is illustrated by curved solid and dashed lines, the two positions occurring at time intervals of one-half period. In the centre of the string is the point at which the string vibrates with its greatest amplitude; this is called an antinodal point, or antinode, and labeled A.

Nodes and antinodes

The next two vibrational modes of the string are also represented in Figure 4. For these vibrations the string is divided into equal segments called loops. Each loop is one-half wavelength long, and the wavelength is related to the length of the string by the following equation:

$$\lambda_n = \frac{2L}{n}. \tag{23}$$

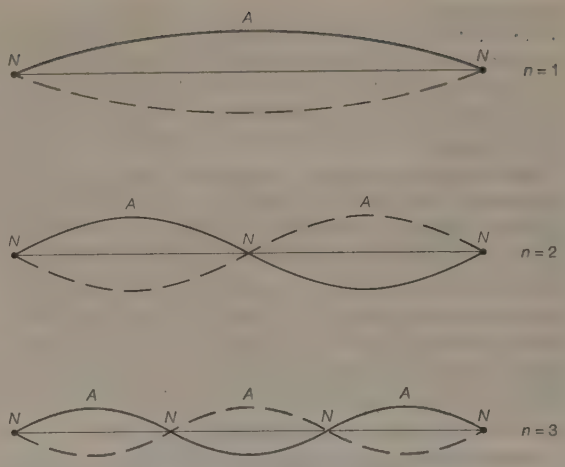


Figure 4: *The first three harmonic standing waves in a stretched string.*
Nodes (N) and antinodes (A) are marked. The harmonic number (n) for each standing wave is given on the right (see text).

Here the integer n equals the number of loops in the standing wave. From equation (22) above, the frequencies of these vibrations (f_n) can be deduced as:

$$f_n = \frac{v}{\lambda_n} = \left(\frac{n}{2L}\right)\sqrt{\frac{F}{\mu}}; \tag{24}$$

or, in terms of the fundamental frequency f_1,

$$f_n = nf_1. \tag{25}$$

Here n is called the harmonic number, because the sequence of frequencies existing as standing waves in the string are integral multiples, or harmonics, of the fundamental frequency.

In the middle representation of Figure 4, labeled $n = 2$ and called the second harmonic, the string vibrates in two sections, so that the string is one full wavelength long. Because the wavelength of the second harmonic is one-half that of the fundamental, its frequency is twice that of the fundamental. Similarly, the frequency of the third harmonic (labeled $n = 3$) is three times that of the fundamental.

Overtones. Another term sometimes applied to these standing waves is overtones. The second harmonic is the first overtone, the third harmonic is the second overtone, and so forth. "Overtone" is a term generally applied to any higher-frequency standing wave, whereas the term harmonic is reserved for those cases in which the frequencies of the overtones are integral multiples of the frequency of the fundamental. Overtones or harmonics are also called resonances. In the phenomenon of resonance, a system that vibrates at some natural frequency is subjected to external vibrations of the same frequency; as a result, the system resonates, or vibrates at a large amplitude.

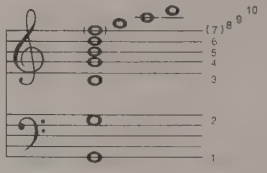


Figure 5: *The first 10 notes in the overtone series of G_2*
The harmonic number of each note is to the right (see text).

The overtone series

The sequence of frequencies defined by equation (25), known as the overtone series, plays an important part in the analysis of musical instruments and musical tone quality. If the fundamental frequency is the note G_2 at the bottom of the bass clef, the first 10 frequencies in the series will correspond closely to the notes shown in Figure 5. Here the frequencies of the octaves (harmonics 1, 2, 4, and 8) are exactly those of the notes shown, but the other frequencies of the overtone series differ by a small

amount from the frequencies of the notes on the equal-tempered scale. The seventh harmonic is quite out of tune when compared with the actual note, so it is enclosed in parentheses.

During the Middle Ages in Europe, keyboard instruments were sometimes tuned to a scale in which the primary chords were true frequencies of the overtone series. This tuning method, called just intonation, provided beatless chords, because the notes in the chord were members of a single overtone series.

Mersenne's laws. From equation (22) can be derived three "laws" detailing how the fundamental frequency of a stretched string depends on the length, tension, and mass per unit length of the string. Known as Mersenne's laws, these can be written as follows:

1. The fundamental frequency of a stretched string is inversely proportional to the length of the string, keeping the tension and the mass per unit length of the string constant:

$$f_1 \propto \frac{1}{L}. \tag{26}$$

2. The fundamental frequency of a stretched string is directly proportional to the square root of the tension in the string, keeping the length and the mass per unit length of the string constant:

$$f_1 \propto \sqrt{F}. \tag{27}$$

3. The fundamental frequency of a stretched string is inversely proportional to the mass per unit length of the string, keeping the length and the tension in the string constant:

$$f_1 \propto \frac{1}{\sqrt{\mu}}. \tag{28}$$

Operation of string instruments

Mersenne's laws help explain the construction and operation of string instruments. The lower strings of a guitar or violin are made with a greater mass per unit length, and the higher strings made thinner and lighter. This means that the tension in all the strings can be made more nearly the same, resulting in a more uniform sound. In a grand piano, the tension in each string is over 100 pounds, creating a total force on the frame of between 40,000 and 60,000 pounds. A large variation in tension between the lower and the higher strings could lead to warping of the piano frame, so that, in order to apply even tension throughout, the higher strings are shorter and smaller in diameter while the bass strings are constructed of heavy wire wound with additional thin wire. This construction makes the wires stiff, causing the overtones to be higher in frequency than the ideal harmonics and leading to the slight inharmonicity that plays an important part in the characteristic piano tone.

In air columns. In a manner analogous to the treatment of standing waves in a stretched string, it is possible to carry out an analysis of the structure of standing waves in air columns. If two identical sinusoidal waves move in opposite directions in a column of air, a standing wave of the same frequency will be formed, just as it is in a string. The standing wave will consist of equally spaced nodes and antinodes with a loop length equal to one-half wavelength in air. Because the motion of the air forming this standing wave is rather complicated, the graphic representation is more abstract, but it can be drawn in a similar manner to that of the string. The simplest standing waves in both open and closed air columns are shown in Figure 6. Each standing wave is identified by its harmonic number (n), and location of the nodes (N) and antinodes (A) are indicated.

Tubes are classified by whether both ends of the tube are open (an open tube) or whether one end is open and one end closed (a closed tube). The basic acoustic difference is that the open end of a tube allows motion of the air; this results in the occurrence there of a velocity or displacement antinode similar to the centre of the fundamental mode of a stretched string, as illustrated at the top of Figure 4. On the other hand, the air at the closed end of a tube cannot move, so that a closed end results in a velocity node similar to the ends of a stretched string.

Open tubes. In an open tube, the standing wave of the lowest possible frequency for that particular length of tube (in other words, the fundamental) has antinodes at each end and a node in the centre. This means that an open tube is one-half wavelength long. The fundamental frequency (f_1) is thus

$$f_1 = \frac{S}{\lambda} = \frac{S}{2L_o}, \tag{29}$$

where L_o is the length of the open tube. The standing wave of each successive harmonic has one additional loop, as shown by $n = 2$ and $n = 3$ in Figure 6. The wavelength (λ_n) of each successive standing wave is calculated as

$$\lambda_n = \frac{2L_o}{n}, \tag{30}$$

and the frequency (f_n) as

$$f_n = \frac{nS}{2L_o} = nf_1, \tag{31}$$

just as in the case of the stretched string.

Closed tubes. The end conditions of a closed tube create a node at the closed end and an antinode at the open end, so that the length of a closed tube (L_c) is one-quarter of a wavelength. For this reason, the length of the closed tubes represented in Figure 6 is one-half that of the open tubes, so that both open and closed tubes produce the same fundamental frequency. In addition, the boundary conditions of a closed tube allow only an odd number of quarter-wavelengths to occupy any given length, so that

$$\lambda = \frac{4L_c}{n}, \tag{32}$$

where only odd n are allowed. Thus, the frequencies of standing waves in a closed tube include only the odd harmonics,

$$f_n = \frac{nS}{4L_c} = nf_1, \tag{33}$$

where values for n are odd integers only.

The Kundt's tube

Measuring techniques. A dramatic device used to "observe" the motion of air in a standing wave is the Kundt's tube. Cork dust is placed on the bottom of this tube, and a standing wave is created. A standing wave in a Kundt's tube consists of a complex series of small cell oscillations, an example of which is illustrated in Figure 7. The air is set in motion, and the vortex motion of the air cells blows the cork dust into small piles, forming a striation pattern. This pattern is very clear and strong at the velocity antinodes of the standing wave, but it disappears at the locations of nodal points. Alternating locations of nodes and antinodes are thus readily observed using this technique.

Under actual conditions, a node is located exactly at the closed end of a tube, but the antinode, owing to the way a wave reflects when it hits the open end, is actually out past the end of the tube by a small distance known as the end correction. The end correction depends primarily

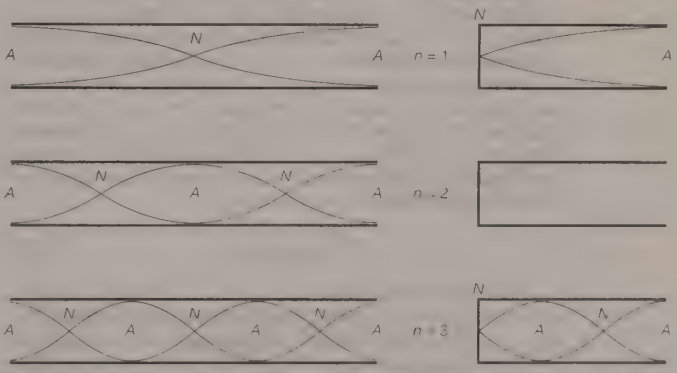


Figure 6: *The first three harmonic standing waves in* (left) *open and* (right) *closed tubes.*
Velocity nodes (N) and antinodes (A) are marked. The harmonic number (n) for each standing wave is given in the centre. The second harmonic does not exist in a closed tube (see text).

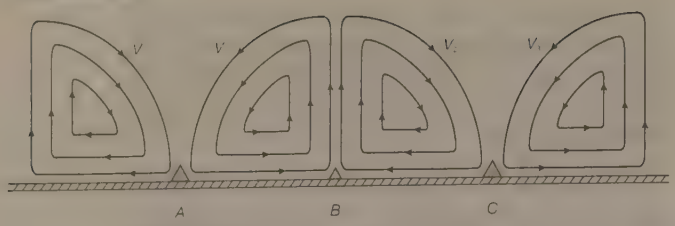


Figure 7: *Motion of air in a standing wave in a Kundt's tube.* In this side view along the tube, cork dust is swept up by vortices V_0–V_4 and deposited at points *A*–*C*, which define the striation patterns of the standing wave (see text).
From R.A. Carman, "Kundt Tube Dust Striations," *American Journal of Physics*, vol. 23 (1955)

on the radius of the tube: it is approximately equal to 0.6 times the radius of an unflanged tube and 0.82 times the radius of a flanged tube. The effective length of the tube, which must be assumed for the value of L in the equations above, incorporates the end correction.

An important feature of this discussion of standing waves in air columns is that the terms node and antinode refer to the places in the vibrating medium where there is zero and maximum displacement or velocity. Many textbooks and reference works use illustrations in which the wave drawn in a tube represents pressure rather than velocity or displacement. In this case, all the nodes and the antinodes are the reverse of those shown in Figure 6—that is, a pressure node (corresponding to a displacement or velocity antinode) occurs at the open end of a tube, while a pressure antinode (corresponding to a displacement or velocity node) occurs at the closed end. Because most microphones respond to changes in pressure, this type of representation may be more useful when discussing experimental observations involving the use of microphones.

In solid rods. A thin metal rod can sustain longitudinal vibrations in much the same way as an air column. The ends of a rod, when free, act as antinodes, while any point at which the rod is held becomes a node, so that the representation of their standing waves is identical to that of an open tube. Such standing waves can be activated by sharply striking the end of the rod with a hard object or by scraping the rod with a cloth or with fingers coated with resin. The harmonic frequencies are then given by

$$f_n = \left(\frac{n}{2L}\right)\sqrt{\frac{Y}{\rho}}, \tag{34}$$

where n is the harmonic number, Y is the Young's modulus (as described above in *Plane waves: The speed of sound*), and ρ is the density of the material. This type of standing wave was used by Ernst Chladni in determining the speed of sound in metals.

In nonharmonic systems. The resonant systems described above have a series of standing-wave resonances that vibrate at the frequencies of the overtone series, but there are several systems whose resonances are not so simply related.

The Helmholtz resonator. An important type of resonator with very different acoustic characteristics is the Helmholtz resonator, named after the German physicist Hermann von Helmholtz. Essentially a hollow sphere with a short, small-diameter neck, a Helmholtz resonator has a single isolated resonant frequency and no other resonances below about 10 times that frequency. The resonant frequency (f) of a classical Helmholtz resonator, shown in Figure 8, is determined by its volume (V) and by the length (L) and area (A) of its neck:

$$f = \left(\frac{S}{2\pi}\right)\sqrt{\frac{A}{LV}}, \tag{35}$$

where S is the speed of sound in air. As with the tubes discussed above, the value of the length of the neck should be given as the effective length, which depends on its radius.

Studying components of musical tones

The isolated resonance of a Helmholtz resonator made it useful for the study of musical tones in the mid-19th century, before electronic analyzers had been invented. When a resonator is held near the source of a sound, the air in it will begin to resonate if the tone being analyzed has a spectral component at the frequency of the resonator. By listening carefully to the tone of a musical instrument with such a resonator, it is possible to identify the spectral components of a complex sound wave such as those generated by musical instruments.

The air cavity of a string instrument, such as the violin or guitar, functions acoustically as a Helmholtz-type resonator, reinforcing frequencies near the bottom of the instrument's range and thereby giving the tone of the instrument more strength in its low range. The acoustic band-pass filter shown in Figure 3D uses a Helmholtz resonator to absorb a band of frequencies from the sound wave passing down an air duct and then reemitting them with the opposite phase, so that they will interfere destructively with the incoming wave and cause it to attenuate. The large jugs used in a jug band also function as Helmholtz resonators, resonating at a single low frequency when air is blown across their openings. Tuning forks are often mounted on boxes, because the air cavity in a box oscillates like a Helmholtz resonator and provides coupling between the tuning fork and the outside air.

Rectangular boxes. An air cavity in the shape of a rectangular box has a sequence of nonharmonic resonances. In such a case the walls are nodal points, and there are standing waves between two parallel walls and mixed standing waves involving several walls. The frequencies of such standing waves are given by the relation

$$f = \left(\frac{S}{2}\right)\sqrt{\left(\frac{N_x}{x}\right)^2 + \left(\frac{N_y}{y}\right)^2 + \left(\frac{N_z}{z}\right)^2}, \tag{36}$$

where x, y, and z are the dimensions of the box and N_x, N_y, and N_z are any integers. In the case where $N_y = N_z = 0$ and $N_x = 1$, the frequency is

$$f = \frac{S}{2x}, \tag{37}$$

corresponding to a half-wavelength the length of the box. This type of resonance is found inside a loudspeaker box, and it must be avoided when tuning a bass reflex speaker port. Such resonances are also readily observed in shower stalls and small rooms such as music practice rooms with parallel walls. Because of these resonances, practice rooms are often made with oblique walls.

Stretched membranes. In a two-dimensional system—for instance, a vibrating plate or a stretched membrane such as a drumhead—the resonant frequencies are not related by integral multiples; that is, their resonances or overtones are inharmonic. Most tuned percussion instruments fall into this category, which is one reason why a tune played on bells or timpani is sometimes more difficult to follow than a tune played on a violin or trumpet. Part of the design goal for tuned bar instruments is to make the shape such that two or more of the resonant frequencies line up like those of wind or string instruments, rendering the pitch clearer. Some, such as the marimba and xylophone, use tubular resonators tuned to the desired frequency of the bar in order to reinforce any overtones that are harmonics of the tube. The South Asian tabla

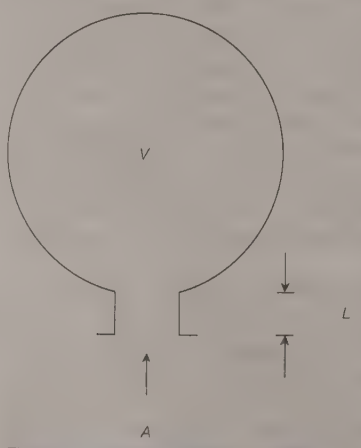


Figure 8: A classic Helmholtz resonator with volume *V* and with a neck of length *L* and cross-sectional area *A*.

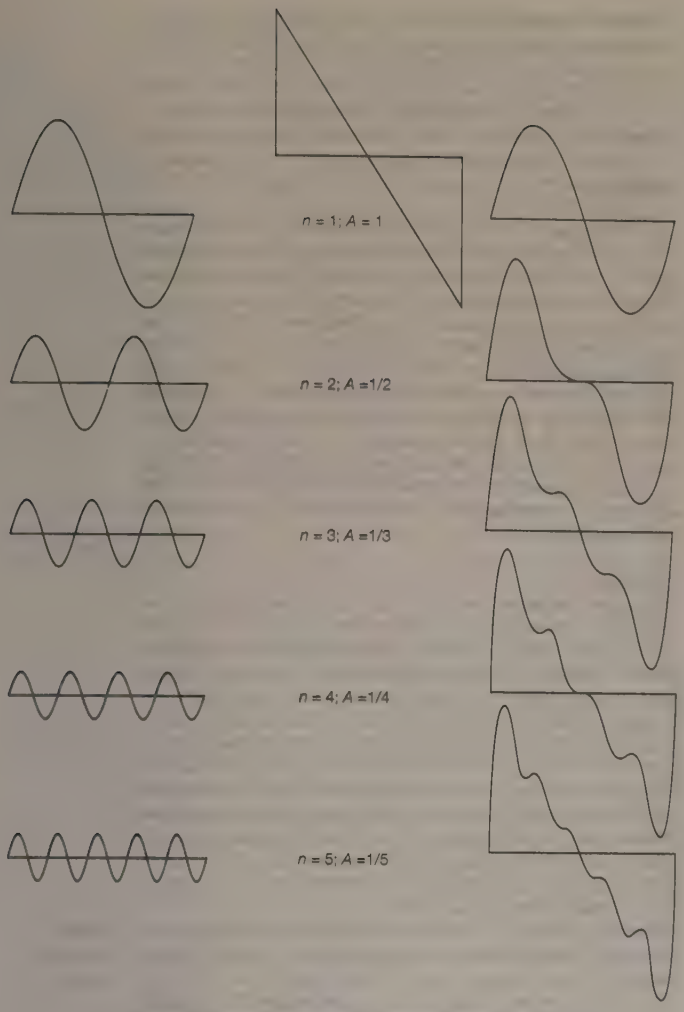


Figure 9: *Fourier synthesis of a complex wave.* The wave at the middle top is synthesized by successively adding to the fundamental each harmonic (n) with an amplitude of $1/n$, as shown at the left. The result is the sawtooth-shaped partial sum waves at the right.

From R.E. Berg and D.G. Stork, *The Physics of Sound* (1982). Prentice-Hall, Inc., Englewood Cliffs, New Jersey

achieves its relatively clear pitch by using a nonuniform, or weighted, drumhead.

STEADY-STATE WAVES

Spectral analysis. *The Fourier theorem.* Fundamental to the analysis of any musical tone is the spectral analysis, or Fourier analysis, of a steady-state wave. According to the Fourier theorem, a steady-state wave is composed of a series of sinusoidal components whose frequencies are those of the fundamental and its harmonics, each component having the proper amplitude and phase. The sequence of components that form this complex wave is called its spectrum.

The sawtooth wave

The synthesis of a complex wave from its spectral components is illustrated by the sawtooth wave in Figure 9. The wave to be synthesized is shown by the graph at the upper middle, with its fundamental to the left and right. Adding the second through fifth harmonics, as shown on the left below the fundamental, results in the sawtooth shapes shown on the right. Figure 10A shows the spectrum of the sawtooth wave taken to the 10th harmonic. In Figure 10B and 10C are the wave form and spectrum of a clarinet playing the note B♭ = 233.08 hertz.

The sound spectrograph. A sound that changes in time, such as a spoken word or a bird call, can be more completely described by examining how the Fourier spectrum changes with time. In a graph called the sound spectrograph, frequency of the complex sound is plotted versus time, with the more intense frequency components shown in the third dimension or more simply as a darker point on a two-dimensional graph. The so-called voiceprint is an example of a sound spectrograph. At one time it was believed that people have voiceprints that are as unique as their fingerprints, so that individuals could be identified by their voiceprints, but the technology of the voiceprint has never been developed. In certain bird atlases, sound spectrographs of bird calls are included with other information, allowing identification of each bird by its call.

Generation by musical instruments. The steady-state tone of any musical instrument can also be analyzed and its Fourier spectrum constructed. The amplitudes of the various spectral components partially determine the tone quality, or timbre, of the instrument.

Bore configuration and harmonicity. The bore shapes of musical instruments, which have developed over the centuries, have rather interesting effects. Cylindrical and conical bores can produce resonances that are harmonics of the fundamental frequencies, but bores that flare faster than a cone create nonharmonic overtones and thus produce raucous tones rather than good musical sounds. A fact discovered by early musical instrument builders, this is the reason why the musical instruments that have developed over the past millennium of Western history are limited to those with either cylindrical or conical bores. In general, a rapidly flaring bell is added to the end of the instrument to reduce the impedance mismatch as the sound emerges from the instrument, thus increasing the ability of the instrument to radiate sound.

Characteristic spectra of musical instruments

The presence of any given harmonic in the spectrum of a particular musical instrument depends on the nature of the vibrating system. For example, if the system functions acoustically as an open tube or a vibrating string, all harmonics will likely be present in the wave. Examples of this are the flute, the recorder, and the violin. On the other hand, the clarinet functions acoustically as a closed tube, because it is cylindrical in shape and has a reed end.

From R.E. Berg and D.G. Stork, *The Physics of Sound* (1982). Prentice-Hall, Inc., Englewood Cliffs, New Jersey

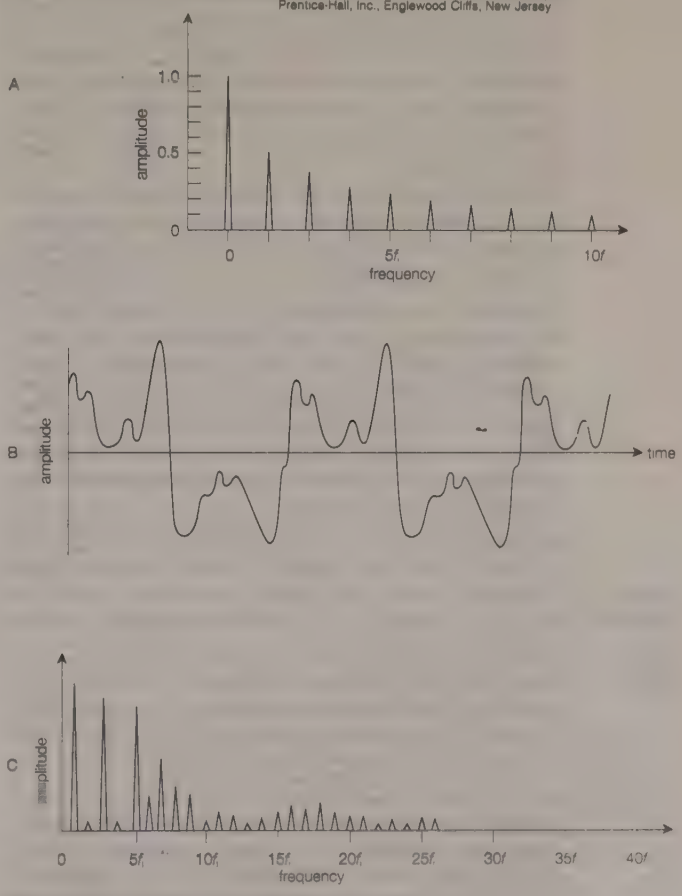


Figure 10: (A) The first 10 components of the Fourier spectrum of a sawtooth wave of frequency f_1. (B) The wave shape and (C) the spectrum of a clarinet playing the note B♭ = 233.08 hertz.

Therefore, as explained above in *Standing waves: In air columns,* the odd harmonics are emphasized in the clarinet spectrum—particularly at low frequencies, as seen in Figure 10C. Other wind instruments function acoustically as open tubes for a variety of reasons. The addition of a mouthpiece and a bell to a tube, either cylindrical or conical, results in all harmonics being possible, as in both the trumpet (cylindrical) and cornet (conical) family of brasses. Even after fixing a reed to one end of a conical tube—as in the oboe, bassoon, and saxophone families—the instruments still function acoustically as open tubes, producing all harmonics. The sawtooth wave, having all harmonics, therefore sounds more like a trumpet or a saxophone than like a clarinet.

Other effects on tone. Because many musical instrument families have similar spectra, there must be other factors that affect their tone quality and by which their tones can be distinguished. Attack transients, such as the way in which a string is bowed, a trumpet tongued, or a piano key struck, and decay transients, such as the way the sound of a plucked string dies away, are very important in many instruments, particularly those that are struck or plucked. Vibrato (a periodic slow change in pitch) and tremolo (a periodic slow change in amplitude) also aid the analysis of steady-state sounds.

Inharmonicities, or deviations of the frequencies of the harmonics from the exact multiples of the fundamental, are very important in tuned percussion instruments. For example, because of the inherent stiffness of piano strings, the overtones of the piano have slight inharmonicities. Indeed, the frequency of the 16th harmonic as played on the piano is about one-half step higher than the exact frequency of the harmonic.

Variations in air pressure. Basic to flutes and recorders, an edge tone is a stream of air that strikes a sharp edge, where it creates pressure changes in the air column that propagate down the tube. Reflections of these pressure variations then force the air stream back and forth across the edge, reinforcing the vibration at the resonant frequency of the tube. The time required to set up this steady-state oscillation is called the transient time of the instrument. The human ear is extremely sensitive to transients in musical tones, and such transients are crucial to the identification of various musical instruments whose spectra are similar.

In musical instruments the pressure variations generated by edge tones, a reed, or the lips set up standing waves in the air column that in turn drive the air stream, reed, or lips. Thus, contrary to common belief, the vibrations of the air column drive the reed or the lips open and closed; the reed or lips do not drive the air column. In the clarinet, for example, air is forced through the reed, creating a pulse of air that travels down the tube. Simultaneously, the reed is pulled closed by pressure of the lips and by rapid air flow out of the reed. After one reflection off the end of the tube, the pulse reflects as a rarefaction, holding the reed shut, but after the second reflection the pulse returns as a compression, forcing the reed open so that the process is repeated.

Acoustic function of the vocal column

The human voice. Groups of emphasized harmonics, known as formants, play a crucial role in the vowel sounds produced by the human voice. Vocal formants arise from resonances in the vocal column. The vocal column is about 17.5 centimetres (7 inches) long, on the average, with its lower end at the vocal folds and its upper end at the lips. Like a reed or like lips at the mouthpiece of a wind instrument, the vocal folds function acoustically as a closed end, so that the vocal column is a closed-tube resonator with resonant frequencies of about 500, 1,500, 2,500, and 3,500 hertz, and so on. The vibration frequency of the vocal folds, determined by the folds' tension, determines the frequency of the vocal sound. When a sound is produced, all harmonics are present in the spectrum, but those near the resonant frequencies of the vocal column are increased in amplitude. These emphasized frequency regions are the vocal formants. By changing the shape of the throat, mouth, and lips, the frequencies of the formants are varied, creating the different vowel sounds.

Noise. The idea of noise is fundamental to the sound of many vibrating systems, and it is useful in describing the spectra of vocal sibilants as well. Just as white light is the combination of all the colours of the rainbow, so white noise can be defined as a combination of equally intense sound waves at all frequencies of the audio spectrum. A characteristic of noise is that it has no periodicity, and so it creates no recognizable musical pitch or tone quality, sounding rather like the static that is heard between stations of an FM radio.

Another type of noise, called pink noise, is a spectrum of frequencies that decrease in intensity at a rate of three decibels per octave. Pink noise is useful for applications of sound and audio systems because many musical and natural sounds have spectra that decrease in intensity at high frequencies by about three decibels per octave. Other forms of coloured noise occur when there is a wide noise spectrum but with an emphasis on some narrow band of frequencies—as in the case of wind whistling through trees or over wires. In another example, as water is poured into a tall cylinder, certain frequencies of the noise created by the gurgling water are resonated by the length of the tube, so that pitch rises as the tube is effectively shortened by the rising water.

HEARING

Dynamic range of the ear. The ear has an enormous range of response, both in frequency and in intensity. The frequency range of human hearing extends over three orders of magnitude, from about 20 hertz to about 20,000 hertz, or 20 kilohertz. The minimum audible pressure amplitude, at the threshold of hearing, is about 10^{-5} pascal, or about 10^{-10} standard atmosphere, corresponding to a minimum intensity of about 10^{-12} watt per square metre. The pressure fluctuation associated with the threshold of pain, meanwhile, is over 10 pascals—one million times the pressure or one trillion times the intensity of the threshold of hearing. In both cases, the enormous dynamic range of the ear dictates that its response to changes in frequency and intensity must be nonlinear.

Equal-loudness curves

Shown in Figure 11 is a set of equal-loudness curves, sometimes called Fletcher-Munson curves after the investigators, the Americans Harvey Fletcher and W.A. Munson, who first measured them. The curves show the varying absolute intensities of a pure tone that has the same loudness to the ear at various frequencies. The determination of each curve, labeled by its loudness level in phons, involves the subjective judgment of a large number of people and is therefore an average statistical result. However, the curves are given a partially objective basis by defining the number of phons for each curve to be the same as the sound intensity level in decibels at 1,000 hertz—a physically measurable quantity. Fletcher and Munson placed the threshold of hearing at 0 phons, or 0 decibels at 1,000

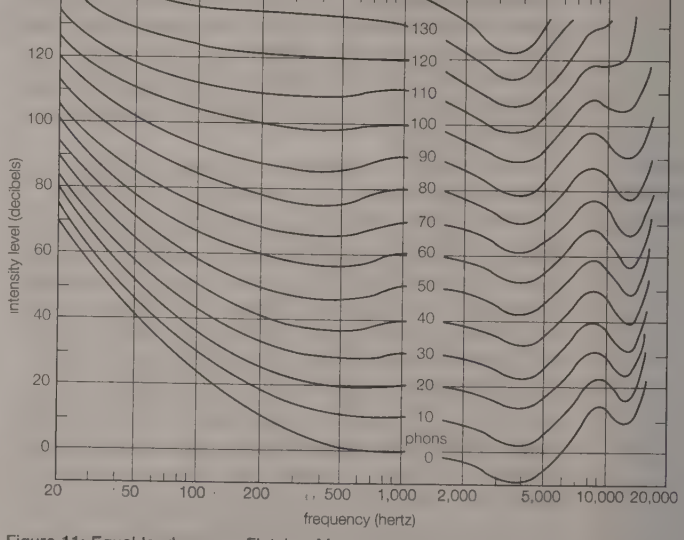


Figure 11: Equal-loudness, or Fletcher-Munson, curves.

hertz, but more accurate measurements now indicate that the threshold of hearing is slightly greater than that. For this reason, the curve labeled 0 phons in Figure 11 is slightly lower than the intensity level of the threshold of hearing over the entire frequency range. The curve labeled 120 phons is sometimes called the threshold of pain, or the threshold of feeling.

Several interesting observations can be made regarding Figure 11. The minimum intensity in the threshold of hearing occurs at about 4,000 hertz. This corresponds to the fundamental frequency at which the ear canal, acting as a closed tube about two centimetres long, has a specific resonance. The pressure variation corresponding to the threshold of hearing, roughly equivalent to placing the wing of a fly on the eardrum, causes a vibration of the eardrum of less than the radius of an atom. If the threshold of hearing did not rise for low frequencies, body sounds, such as heartbeat and blood pulsing, would be continually audible. Music is normally played at intensity levels between about 30 and 100 decibels. When it is played more softly, decreasing the sound level of all frequencies by the same amount, bass frequencies fall below the threshold of hearing. This is why the loudness control on an audio system raises the intensity of low frequencies—so that the music will have the same proportion of treble and bass to the ear as when it is played at a higher level.

As stated above, the ear has an enormous dynamic range, the threshold of pain corresponding to an intensity 12 orders of magnitude (10^{12} times) greater than the threshold of hearing. This leads to the necessity of a nonlinear intensity response. In order to be sensitive to intense waves and yet remain sensitive to very low intensities, the ear must respond proportionally less to higher intensity than to lower intensity. This response is logarithmic, because the ear responds to ratios rather than absolute pressure or intensity changes. At almost any region of the Fletcher-Munson diagram, the smallest change in intensity of a sinusoidal sound wave that can be observed, called the intensity just noticeable difference, is about one decibel (further reinforcing the value of the decibel intensity scale). One decibel corresponds to an absolute energy variation of a factor of about 1.25. Thus, the minimum observable change in the intensity of a sound wave is greater by a factor of nearly 10^{12} at high intensities than it is at low intensities.

Frequency response of the ear

The frequency response of the ear is likewise nonlinear. Relating frequency to pitch as perceived by the musician, two notes will "sound" similar if they are spaced apart in frequency by a factor of two, or octave. This means that the frequency interval between 100 and 200 hertz sounds the same as that between 1,000 and 2,000 hertz or between 5,000 and 10,000 hertz. In other words, the tuning of musical scales and musical intervals is associated with frequency ratios rather than absolute frequency differences in hertz. As a result of this empirical observation that all octaves sound the same to the ear, each frequency interval equivalent to an octave on the horizontal axis of the Fletcher-Munson scale is equal in length.

The audio frequency range encompasses nearly nine octaves. Over most of this range, the minimum change in the frequency of a sinusoidal tone that can be detected by the ear, called the frequency just noticeable difference, is about 0.5 percent of the frequency of the tone, or about one-tenth of a musical half-step. The ear is less sensitive near the upper and lower ends of the audible spectrum, so that the just noticeable difference becomes somewhat larger.

The ear as spectrum analyzer. The ear actually functions as a type of Fourier analysis device, with the mechanism of the inner ear converting mechanical waves into electrical impulses that describe the intensity of the sound as a function of frequency. Ohm's law of hearing is a statement of the fact that the perception of the tone of a sound is a function of the amplitudes of the harmonics and not of the phase relationships between them. This is consistent with the place theory of hearing, which correlates the observed pitch with the position along the basilar membrane of the inner ear that is stimulated by the corresponding frequency.

The intensity level at which a sound can be heard is affected by the existence of other stimuli. This effect, called masking, plays an important role in the psychophysical response to sound. Low frequencies mask higher frequencies much more strongly than high frequencies mask lower ones; this is one reason why a complex wave is perceived as having a different tone quality or timbre from a pure wave of the same frequency, even though they have the same pitch. Noise of low frequencies can be used to mask unwanted distracting sounds, such as nearby conversation in an office, and to create greater privacy.

Masking

The ear is responsive to the periodicity of a wave, so that it will hear the frequency of a complex wave as that of the fundamental whether or not the fundamental is actually present as a component in the wave, although the wave will have a different timbre than it would were the fundamental actually present. This effect, known as the missing fundamental, subjective fundamental, or periodicity pitch, is used by the ear to create the fundamental in sound radiating from a small loudspeaker that is not capable of providing low frequencies.

If the intensity of a sound is sufficiently great, the wave shape will be distorted by the ear mechanism, owing to its nonlinearity. The spectral analysis of the sound will then include frequencies that are not present in the sound wave, causing a distorted perception of the sound. If two or more sounds of great intensity are presented to the ear, this effect will introduce what are called combination tones. Two pure tones of frequency f_1 and f_2 will create a series of new pure tones: the sum tones,

$$f_+ = nf_1 + mf_2; \tag{38}$$

and the difference tones,

$$f_- = |nf_1 - mf_2|. \tag{39}$$

(Here n and m are any two integers.) Sum tones are difficult to hear because they are masked by the higher-intensity tones creating them, but difference tones are often observed in musical performance. For example, if the two tones are adjacent members of the harmonic series, the fundamental of that series will be produced as a difference tone, enhancing the ability of the ear to identify the fundamental pitch.

Binaural perception. The paths from the ears to the brain are separate; that is, each ear converts the sound reaching it into electrical impulses, so that sounds from the two ears mix in the brain not as physical vibrations but as electrical signals. This separation of pathways has the direct result that, if two pure tones are presented to each ear separately (*i.e.*, binaurally) at low levels, it will be very difficult for the ears to compare the frequencies because with no direct mixing of the mechanical waves there will be no regular beats. This difference in pitch perception between the two ears, called diplacusis, is generally not a problem. A type of beating known as binaural beats can sometimes be observed when the two tones are presented binaurally.

Also, two tones very close to an octave apart produce another type of monaural beating as they change in phase. This effect, known as second-order beats or quality beats, is observed as a slight periodic change in the quality of the combined tone. It serves as a counterexample to Ohm's law of hearing, which suggests that the quality of a sound depends only on the amplitudes of the harmonics and not on their phases.

Although the two ears are not connected by mechanical means, the brain is sensitive to phase and is able to determine the phase relationship between stimuli presented to the two ears. Locating a sound source laterally in space makes use of fundamental properties of sound waves as well as the ability of the brain to identify the phase difference between signals from the two ears. At low frequencies, where the wavelength is large and the waves diffract strongly, the brain is able to perceive the phase difference between the same sound reaching both ears, and it can thus locate the direction from which the sound is coming. On the other hand, at high frequencies the wavelength may be so short that there may be more than one period of time delay between the signals arriving at the two ears,

Locating sound by phase difference

creating an ambiguity in the phase difference. Fortunately, at these high frequencies there is so much less diffraction of sound waves that the head actually shields one ear more than the other. In such cases the difference in intensity of the sound waves reaching the two ears, rather than their phase difference, is used by the ears in spatial localization. Spatial localization in the vertical direction is poor for most people.

Applications of sound

ELECTROMECHANICAL TRANSDUCERS

Basic to the operation of the microphone and the loudspeaker is the electroacoustical transducer. A transducer is any type of electromechanical device that either converts an electrical signal into sound waves (as in a loudspeaker) or converts a sound wave into an electrical signal (as in the microphone). Many of the transducers used in everyday life operate in both directions, such as the speakerphone on certain intercoms.

Microphones. In order to evaluate the adequacy of various types of microphones for specific uses, one must consider the linearity of their frequency response, their directional characteristics, their durability, and their cost.

Linearity and directivity. Frequency linearity is the ability of a microphone to yield an electrical output that is proportional to the amplitude of the sound input over the entire frequency range. For music, this must extend to much lower and much higher frequencies than for voice use only. For high-quality musical reproduction, the flatter, or more nearly linear, microphone is generally the best choice.

Microphones also have directional characteristics. Those that uniformly pick up signals coming from all directions are referred to as omnidirectional. A common directional microphone is the cardioid microphone, so called because, when the intensity response as a function of angle is plotted on a polar graph, the curve is heart-shaped. A cardioid microphone is useful for recording live performances, where it is desirable to eliminate audience noise. A shotgun microphone has a very strong forward directional response. A parabolic reflector, similar to that of a reflecting telescope, is used to pick up and amplify relatively weak sounds coming from a certain direction. This is useful for such diverse applications as listening to bird calls and listening to the quarterback's signals at a football game.

Types of transducers. Most microphones use either an electromagnetic or an electrostatic technique to convert sound waves into electrical signals. The dynamic microphone is constructed with a small magnet that oscillates inside a coil attached to the diaphragm. When a sound wave causes the diaphragm of the microphone to vibrate, the relative motion of the magnet and coil creates an electrical signal by magnetic induction. Either a moving-coil or a moving-magnet system may be employed, depending on which element is connected to the moving diaphragm; the moving coil is used more often. The dynamic microphone is rugged and has reasonably good linearity, so that high-quality models are useful in recording. Because a moving-coil microphone and a moving-coil loudspeaker are very similar, intercoms are often made with the same element serving both functions.

The dynamic microphone

The electrostatic or condenser microphone is constructed with the diaphragm as one plate of a parallel-plate capacitor. The most popular form of this type of microphone is the electret condenser microphone, in which the plates are given a permanent electrical charge. When a sound wave causes the charged diaphragm plate to vibrate, the voltage across the plates changes, creating a signal that can be amplified and transmitted to the recording device. An amplifier is often mounted in the microphone, so this type of microphone requires the use of a battery to power the amplifier. Because the diaphragm of a condenser microphone can be very light, compared with the more massive dynamic microphone, it is able to respond faster and at higher frequencies. Consequently, condenser microphones generally have better linearity and a greater frequency range than dynamic microphones.

The crystal microphone uses a piezoelectric crystal as its transducer. Piezoelectric crystals are durable and cheap, and they have relatively large electrical output; for this reason, they are often used in telephones and portable sound systems. They do not have very good linearity and so are inadequate for quality sound recording.

The ribbon microphone is unique in that it responds to the air velocity of the sound wave, not to the pressure variation. Because ribbon microphones are very sensitive, they cannot be used where they will suffer mechanical shocks. Ribbon microphones are bidirectional and can be used to pick up sounds coming from both sides of the microphone equally well.

Loudspeakers. *Electromagnetic speakers.* Most loudspeakers are of the electromagnetic, or dynamic, variety, in which a voice coil moves in the gap of a permanent magnet when a time-varying current flows through the coil. The magnet is generally in the shape of a "W" or a ring. The diaphragm, or cone, of such a loudspeaker moves with the coil, converting the electric current in the coil into a pressure wave. A lit candle placed in front of a loudspeaker cone that is oscillating at about 10 hertz can render the sound wave "visible," as the flame vibrates back and forth longitudinally with the air.

As is the case with microphones, loudspeakers are evaluated largely on their frequency linearity. In order to achieve good frequency response at low frequencies, it is necessary to use a rather large cone; however, owing to the relatively large mass of the loudspeaker coil and cone, it is difficult to achieve good response at high frequencies with the same loudspeaker. Response can be improved by using rather large magnets, but these make a good loudspeaker rather heavy. In addition, the suspension of the coil in the magnet gap is critical, because it must provide for both rapid response and quick damping to its equilibrium position when the signal ceases. Each loudspeaker has a frequency at which it resonates most readily. For large loudspeakers this resonant frequency is useful in enhancing the bass response of the system.

Loudspeaker enclosures

Loudspeakers are mounted in a box, horn, or other enclosure in order to separate the waves from the front and the rear of the loudspeaker and thereby prevent them from canceling each other. (This type of destructive interference is discussed above in *Circular and spherical waves: Interference.*) The most common type of enclosure is the acoustic suspension system, in which the loudspeaker is mounted in an airtight box. To prevent resonances in the box of the type described by equation (36), the inside is generally coated with some sound-absorbent material. Because of the airtight seal, the cone must compress and expand the air inside the box as it moves, so that this type of system is not very efficient in converting electrical energy into sound, especially at bass frequencies.

The tuned port or bass reflex enclosure achieves greater efficiency and extends the bass frequency range by carefully adjusting the shape and position of a hole or tube connecting the inside of the speaker box with the outside. The volume of the box thus acts as a type of Helmholtz resonator, with a resonant frequency that is determined by the geometry of the hole or tube and is deliberately chosen so that it extends the frequency range of the speaker system smoothly to a significantly lower value. In addition, the existence of the port greatly reduces the air pressure variation inside the box, allowing the loudspeaker cone to move much more freely. For these reasons, the typical bass reflex enclosure is much more efficient than the typical acoustic suspension system.

A horn enclosure uses a flared tube to obtain the best acoustic coupling between the loudspeaker cone and the outside, thereby radiating the best possible coherent wave from the speaker cone. Such a system is extremely efficient and is therefore used in public-address systems, open-air theatres, or other places in which great acoustic power is desired. Because a good quality bass horn enclosure is very large, such devices often use bent or folded tubes. The Klipschorn, named for its inventor, the American engineer Paul W. Klipsch, uses the walls in the corner of a room as part of the flared horn.

Because high efficiency and linearity of a single speaker

cannot be extended over the entire audible frequency range, loudspeaker systems are generally formed from two or more individual loudspeakers. A larger speaker, or woofer, produces the lower frequencies, while a smaller speaker, or tweeter, produces the higher frequencies. In such a two-way system, a passive electronic circuit called a crossover network is employed to direct the higher and lower frequencies to the appropriate loudspeaker. A larger or more efficient three-way system may add a midrange speaker, helping to create a more nearly linear response between woofer and tweeter.

Loudspeakers in large areas

The loudspeaker arrays regularly seen in large auditoriums often make use of a single woofer and a single midrange speaker but two or even three high-frequency tweeters. The necessity for using a greater number of tweeters arises from the relatively smaller diffraction of high-frequency (or low-wavelength) sound waves. Because these spread out less and are therefore more directional, it may be necessary to provide several tweeters and aim them so as to cover the entire auditorium. This is unnecessary for the woofer because of the large diffraction of long wavelengths.

Electrostatic speakers. Electrostatic loudspeakers make use of a large, thin metal plate between two parallel screens. An amplified audio signal is impressed onto the screens, polarizing the metal sheet, and the resulting electrostatic force creates a motion of the sheet, producing a sound wave. Electrostatic speakers function well at high frequencies, but they are unable to move enough air to perform well at low frequencies and often require somewhat greater power than electromagnetic speakers. Because of these limitations and other technical problems, they have seen only limited use and are not popular in consumer audio systems.

SOUND RECORDING

The three most popular recording formats used in home audio systems are the long-playing phonograph disc, the audiocassette, and the compact disc.

The phonograph disc. A monaural phonograph record makes use of a spiral 90° V-shaped groove impressed into a plastic disc. As the record revolves at 33⅓ rotations per minute, a tiny "needle," or stylus, simultaneously moves along the groove and vibrates back and forth parallel to the surface of the disc and perpendicular to the groove, tracing out the sound wave. The upper end of the stylus is connected to a tiny magnet, which moves back and forth through a small coil, inducing an electrical voltage that recreates the recorded sound wave. The rate of oscillation of the stylus determines the frequency of the sound, while the amplitude of the oscillation determines its loudness.

Just as the use of two eyes creates a perception of depth, so can the effect of musical "presence" be achieved by stereophonics, recording music with two appropriately positioned microphones and playing it back on two separated loudspeakers. A stereophonic recording provides the two separate signal channels as oscillations perpendicular to either one or the other of the faces of the record groove. The single coil of the monaural pickup is replaced by two coils, which sense the motion of the stylus perpendicular to each groove wall; the inside wall is used as the left channel and the outside wall as the right channel. These two signals are then fed into an audio amplifier and to the loudspeakers.

Limiting variation in frequency

The criterion for frequency control of a recording is that the variation in frequency should not be observable to the ear—*i.e.*, less than about 0.1 percent, which is less than the just noticeable difference in frequency over most of the audible frequency range. In order to eliminate both slow variations in pitch of the recording, called wow, and rapid variations, called flutter, the rotation speed of the record is carefully controlled by use of a heavy turntable and a precision motor. Mechanical vibration of the turntable is isolated from the stylus to avoid "rumble." The stylus itself is elliptical in shape, with the long axis of the ellipse oriented across the groove. In order to achieve good compliance—that is, the ability of the stylus to track the groove and produce a linear signal—the tip of the stylus must be less than 25 micrometres (0.025 millimetre, or $^1/_{1000}$ inch) in size, so that it is generally made of industrial diamond.

Faraday's law of magnetic induction introduces some important features into the science of phonograph records. According to this law, the electric potential induced in the coil of the magnetic pickup is directly proportional to the magnetic field of the moving magnet in the pickup and inversely proportional to the period of the oscillation. This means that, in order to produce a sound wave of constant amplitude at all frequencies, it is necessary to reduce the amplitude of the motion of the stylus at high frequencies and greatly increase that motion at low frequencies. Unfortunately, limitations in the compliance of such a recording system make it impossible for the stylus to accurately track a sufficiently large-amplitude oscillation at low frequencies. Furthermore, the inherent graininess of the plastic from which phonographic recordings are pressed creates high-frequency vibrations of the stylus, which are heard as high-frequency noise, or hiss. Because of these problems, the electrical signal must be amplified at very high frequencies, attenuated at very low frequencies, and approximately linearized for midrange frequencies before the signal is converted into a groove shape and impressed onto the plastic disc—a process called pre-emphasis. Upon playback this sequence is reversed in a process called equalization, providing the listener with a linear and realistic sound.

The audiotape. Audiocassette tape recording also makes use of electromagnetic phenomena to record and reproduce sound waves. The tape consists of a plastic backing coated with a thin layer of tiny particles of magnetic powder, usually ferric oxide (Fe_2O_3) and to a lesser extent chromium dioxide (CrO_2). The recording head of the tape deck consists of a tiny C-shaped magnet with its gap adjacent to the moving tape. The incoming sound wave, having been converted by a microphone into an electrical signal, produces a time-varying magnetic field in the gap of the magnet. As the tape moves past the recording head the powder is magnetized in such a way that the tape carries a record of the shape of the wave being recorded. The frequency of the impressed signal determines the distance along the tape over which the impressed magnetic field must be reversed, and the amplitude of the signal determines the extent of the magnetization of the tape.

Problems in magnetic recording

There are inherent problems with the magnetic recording system. As magnetic domains are flipped in magnetizing the material, they exhibit a certain magnetic inertia, or unwillingness to respond, so that it requires a greater magnetic field than expected to magnetize the oxide on the tape. This effect, known as hysteresis, leads to distortion of the wave shape on the tape. In order to overcome this problem, a sinusoidal signal of about 100 kilohertz is added to the wave immediately before the wave is impressed onto the tape. Known as equalization bias, this signal has the effect of linearizing an inherently nonlinear magnetic medium, largely eliminating distortion.

Another problem arises from the inability of the recording system to organize completely the magnetic domains in these tiny magnetic crystals. The resulting random orientation of the domains results in random noise, which is heard by the listener as tape hiss. Because lower frequencies are more effective in magnetizing the tape, and because the random variation in magnetization is a microscopic effect, tape hiss is primarily a high-frequency phenomenon. Several systems have been designed to deal with this problem, the most prevalent of which is Dolby noise reduction. In the Dolby system the higher-frequency components of a sound wave are amplified before the signal is impressed on the tape so that their amplitudes are well above the amplitude of the tape hiss. On playback, the high frequencies are attenuated after they are read off the tape, reducing their amplitudes to the correct level.

The compact disc. The compact disc, or digital disc, uses digital technology to avoid or mitigate some of the technical problems and requirements inherent in phonograph and audiotape recording. Whereas both phonograph recordings and audiotape have limited dynamic range and frequency response, the compact disc has both a greater dynamic range—ideally, over 90 decibels—and a linear

frequency response from less than 20 hertz to over 20,000 hertz—greater than that of the human ear.

Digital recording uses sampling of the sound wave at a series of points at equal time intervals along the wave to approximate the full wave. In order to maintain frequency response up to 20 kilohertz, the limit of human hearing, it is necessary to sample at slightly above twice that frequency, so that compact discs actually have a sample rate of 44.1 kilohertz. The signal level is divided into 2^{15} (about 32,000) equal intervals. With such a large number of intervals being employed, both large and small wave intensities can be reproduced accurately. Indeed, intensity variations of less than one decibel (the approximate value of the intensity just noticeable difference of the ear) can be achieved over the entire dynamic range of the compact disc.

Accurate reproduction of wave intensity

Each sampled point on the wave is encoded in binary form, and a series of points are impressed on the compact disc. Playback is essentially the reverse of recording. Each point on the wave is read in and stored in a computer memory called a first-in first-out buffer. Using an internal 44.1-kilohertz clock, each point is converted in order into analog form and then input into a standard power amplifier and loudspeaker. The time scale for the recording is exactly reproduced, eliminating the frequency instabilities inherent in other types of recording.

ARCHITECTURAL ACOUSTICS

Reverberation time. Although architectural acoustics has been an integral part of the design of structures for at least 2,000 years, the subject was only placed on a firm scientific basis at the beginning of the 20th century by Wallace Sabine. Sabine pointed out that the most important quantity in determining the acoustic suitability of a room for a particular use is its reverberation time, and he provided a scientific basis by which the reverberation time can be determined or predicted.

When a source creates a sound wave in a room or auditorium, observers hear not only the sound wave propagating directly from the source but also the myriad reflections from the walls, floor, and ceiling. These latter form the reflected wave, or reverberant sound. After the source ceases, the reverberant sound can be heard for some time as it grows softer. The time required, after the sound source ceases, for the absolute intensity to drop by a factor of 10^6—or, equivalently, the time for the intensity level to drop by 60 decibels—is defined as the reverberation time (RT, sometimes referred to as RT_{60}). Sabine recognized that the reverberation time of an auditorium is related to the volume of the auditorium and to the ability of the walls, ceiling, floor, and contents of the room to absorb sound. Using these assumptions, he set forth the empirical relationship through which the reverberation time could be determined:

$$RT = \frac{0.05V}{A}, \tag{40}$$

where RT is the reverberation time in seconds, V is the volume of the room in cubic feet, and A is the total sound absorption of the room, measured by the unit sabin. The sabin is the absorption equivalent to one square foot of perfectly absorbing surface—for example, a one-square-foot hole in a wall or five square feet of surface that absorbs 20 percent of the sound striking it.

The sabin

Both the design and the analysis of room acoustics begin with equation (40). Using this equation and the absorption coefficients of the materials from which the walls are to be constructed, an approximation can be obtained for the way in which the room will function acoustically.

Table 6: Absorption Coefficients of Common Materials at Several Frequencies

	frequency (hertz)					
material	125	250	500	1,000	2,000	4,000
Concrete	0.01	0.01	0.02	0.02	0.02	0.03
Plasterboard	0.20	0.15	0.10	0.08	0.04	0.02
Acoustic board	0.25	0.45	0.80	0.90	0.90	0.90
Curtains	0.05	0.12	0.25	0.35	0.40	0.45

Absorbers and reflectors, or some combination of the two, can then be used to modify the reverberation time and its frequency dependence, thereby achieving the most desirable characteristics for specific uses. Representative absorption coefficients—showing the fraction of the wave, as a function of frequency, that is absorbed when a sound hits various materials—are given in Table 6. The absorption from all the surfaces in the room are added together to obtain the total absorption (A).

While there is no exact value of reverberation time that can be called ideal, there is a range of values deemed to be appropriate for each application. These vary with the size of the room, but the averages can be calculated and indicated by lines on a graph, as in Figure 12. The need for clarity in understanding speech dictates that rooms used for talking must have a reasonably short reverberation time. On the other hand, the full sound desirable in the performance of music of the Romantic era, such as Wagner operas or Mahler symphonies, requires a long reverberation time. Obtaining a clarity suitable for the light, rapid passages of Bach or Mozart requires an intermediate value of reverberation time. For playing back recordings on an audio system, the reverberation time should be short, so as not to create confusion with the reverberation time of the music in the hall where it was recorded.

From R.E. Berg and D.G. Stork, *The Physics of Sound* (1982), Prentice-Hall, Inc., Englewood Cliffs, New Jersey

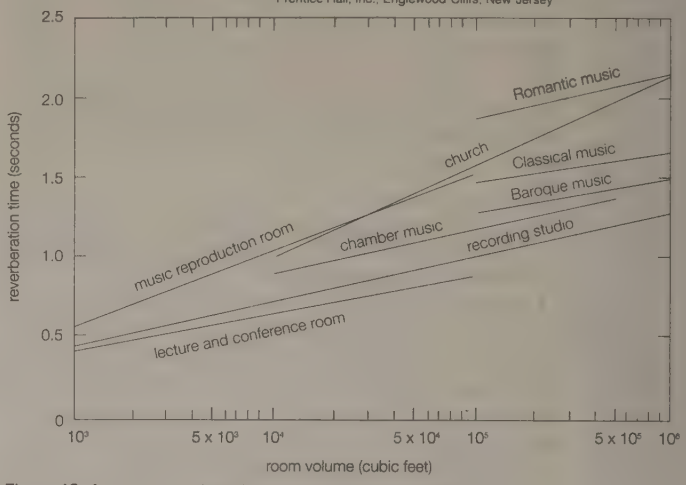


Figure 12: Average reverberation time versus room volume required to achieve optimal response for various types of room and music.

Acoustic criteria. Many of the acoustic characteristics of rooms and auditoriums can be directly attributed to specific physically measurable properties. Because the music critic or performing artist uses a different vocabulary to describe these characteristics than does the physicist, it is helpful to survey some of the more important features of acoustics and correlate the two sets of descriptions.

"Liveness" refers directly to reverberation time. A live room has a long reverberation time and a dead room a short reverberation time. "Intimacy" refers to the feeling that listeners have of being physically close to the performing group. A room is generally judged intimate when the first reverberant sound reaches the listener within about 20 milliseconds of the direct sound. This condition is met easily in a small room, but it can also be achieved in large halls by the use of orchestral shells that partially enclose the performers. Another example is a canopy placed above a speaker in a large room such as a cathedral: this leads to both a strong and a quick first reverberation and thus to a sense of intimacy with the person speaking.

The amplitude of the reverberant sound relative to the direct sound is referred to as fullness. Clarity, the opposite of fullness, is achieved by reducing the amplitude of the reverberant sound. Fullness generally implies a long reverberation time, while clarity implies a shorter reverberation time. A fuller sound is generally required of Romantic music or performances by larger groups, while more clarity would be desirable in the performance of rapid passages from Bach or Mozart or in speech.

Fullness and clarity

"Warmth" and "brilliance" refer to the reverberation time at low frequencies relative to that at higher frequencies. Above about 500 hertz, the reverberation time should be the same for all frequencies. But at low frequencies an increase in the reverberation time creates a warm sound, while, if the reverberation time increased less at low frequencies, the room would be characterized as more brilliant.

"Texture" refers to the time interval between the arrival of the direct sound and the arrival of the first few reverberations. To obtain good texture, it is necessary that the first five reflections arrive at the observer within about 60 milliseconds of the direct sound. An important corollary to this requirement is that the intensity of the reverberations should decrease monotonically; there should be no unusually large late reflections.

"Blend" refers to the mixing of sounds from all the performers and their uniform distribution to the listeners. To achieve proper blend it is often necessary to place a collection of reflectors on the stage that distribute the sound randomly to all points in the audience.

Although the above features of auditorium acoustics apply to listeners, the idea of ensemble applies primarily to performers. In order to perform coherently, members of the ensemble must be able to hear one another. Reverberant sound cannot be heard by the members of an orchestra, for example, if the stage is too wide, has too high a ceiling, or has too much sound absorption on its sides.

Acoustic problems. Certain acoustic problems often result from improper design or from construction limitations. If large echoes are to be avoided, focusing of the sound wave must be avoided. Smooth, curved reflecting surfaces such as domes and curved walls act as focusing elements, creating large echoes and leading to bad texture. Improper blend results if sound from one part of the ensemble is focused to one section of the audience. In addition, parallel walls in an auditorium reflect sound back and forth, creating a rapid, repetitive pulsing of sound known as flutter echo and even leading to destructive interference of the sound wave. Resonances at certain frequencies should also be avoided by use of oblique walls.

Acoustic shadows

Acoustic shadows, regions in which some frequency regions of sound are attenuated, can be caused by diffraction effects as the sound wave passes around large pillars and corners or underneath a low balcony. Large reflectors called clouds, suspended over the performers, can be of such a size as to reflect certain frequency regions while allowing others to pass, thus affecting the mixture of the sound.

External noise can be a serious problem for halls in urban areas or near airports or highways. One technique often used for avoiding external noise is to construct the auditorium as a smaller room within a larger room. Noise from air blowers or other mechanical vibrations can be reduced using the techniques discussed above (see *Circular and spherical waves: Impedance*) and by isolating air handlers.

Good acoustic design must take account of all these possible problems while emphasizing the desired acoustic features. One of the problems in a large auditorium involves simply delivering an adequate amount of sound to the rear of the hall. The intensity of a spherical sound wave decreases in intensity at a rate of six decibels for each factor of two increase in distance from the source, as shown above (see *Circular and spherical waves: Attenuation*). If the auditorium is flat, a hemispherical wave will result. Absorption of the diffracted wave by the floor or audience near the bottom of the hemisphere will result in even greater absorption, so that the resulting intensity level will fall off at twice the theoretical rate, at about 12 decibels for each factor of two in distance. Because of this absorption, the floors of an auditorium are generally sloped upward toward the rear.

ULTRASONICS

The term ultrasound refers to vibrations of frequencies greater than the upper limit of the audible range for humans—that is, greater than about 20 kilohertz. The term sonic is applied to ultrasound waves of very high amplitudes. Hypersound, sometimes called praetersound or microsound, is sound waves of frequencies greater than 10^{13} hertz. At such high frequencies it is very difficult for a sound wave to propagate efficiently; indeed, above a frequency of about 1.25×10^{13} hertz, it is impossible for longitudinal waves to propagate at all, even in a liquid or a solid, because the molecules of the material in which the waves are traveling cannot pass the vibration along rapidly enough.

Many animals have the ability to hear sounds in the human ultrasonic frequency range. Some ranges of hearing for mammals and insects are compared with those of humans in Table 7. A presumed sensitivity of roaches and rodents to frequencies in the 40 kilohertz region has led to the manufacture of "pest controllers" that emit loud sounds in that frequency range to drive the pests away, but they do not appear to work as advertised.

Table 7: Frequency Range of Hearing for Humans and Selected Animals

animal	frequency (hertz)	
	low	high
Humans	20	20,000
Cats	100	32,000
Dogs	40	46,000
Horses	31	40,000
Elephants	16	12,000
Cattle	16	40,000
Bats	1,000	150,000
Grasshoppers and locusts	100	50,000
Rodents	1,000	100,000
Whales and dolphins	70	150,000
Seals and sea lions	200	55,000

Transducers. An ultrasonic transducer is a device used to convert some other type of energy into an ultrasonic vibration. There are several basic types, classified by the energy source and by the medium into which the waves are being generated. Mechanical devices include gas-driven, or pneumatic, transducers such as whistles as well as liquid-driven transducers such as hydrodynamic oscillators and vibrating blades. These devices, limited to low ultrasonic frequencies, have a number of industrial applications, including drying, ultrasonic cleaning, and injection of fuel oil into burners. Electromechanical transducers are far more versatile and include piezoelectric and magnetostrictive devices. A magnetostrictive transducer makes use of a type of magnetic material in which an applied oscillating magnetic field squeezes the atoms of the material together, creating a periodic change in the length of the material and thus producing a high-frequency mechanical vibration. Magnetostrictive transducers are used primarily in the lower frequency ranges and are common in ultrasonic cleaners and ultrasonic machining applications.

Piezoelectric transducers

By far the most popular and versatile type of ultrasonic transducer is the piezoelectric crystal, which converts an oscillating electric field applied to the crystal into a mechanical vibration. Piezoelectric crystals include quartz, Rochelle salt, and certain types of ceramic. Piezoelectric transducers are readily employed over the entire frequency range and at all output levels. Particular shapes can be chosen for particular applications. For example, a disc shape provides a plane ultrasonic wave, while curving the radiating surface in a slightly concave or bowl shape creates an ultrasonic wave that will focus at a specific point.

Piezoelectric and magnetostrictive transducers also are employed as ultrasonic receivers, picking up an ultrasonic vibration and converting it into an electrical oscillation.

Applications in research. One of the important areas of scientific study in which ultrasonics has had an enormous impact is cavitation. When water is boiled, bubbles form at the bottom of the container, rise in the water, and then collapse, leading to the sound of the boiling water. The boiling process and the resulting sounds have intrigued people since they were first observed, and they were the object of considerable research and calculation by the British physicists Osborne Reynolds and Lord Rayleigh, who applied the term cavitation to the process of formation of bubbles. Because an ultrasonic wave can be used carefully to control cavitation, ultrasound has been a

useful tool in the investigation of the process. The study of cavitation has also provided important information on intermolecular forces.

Research is still being carried out on aspects of the cavitation process and its applications. A contemporary subject of research involves emission of light as the cavity produced by a high-intensity ultrasonic wave collapses. This effect, called sonoluminescence, is believed to create instantaneous temperatures hotter than the surface of the Sun.

The speed of propagation of an ultrasonic wave is strongly dependent on the viscosity of the medium. This property can be a useful tool in investigating the viscosity of materials. Because the various parts of a living cell are distinguished by differing viscosities, acoustical microscopy can make use of this property of cells to "see" into living cells, as will be discussed below in *Medical applications.*

Ranging and navigating. Sonar (sound navigation and ranging) has extensive marine applications. By sending out pulses of sound or ultrasound and measuring the time required for the pulses to reflect off a distant object and return to the source, the location of that object can be ascertained and its motion tracked. This technique is used extensively to locate and track submarines at sea and to locate explosive mines below the surface of the water. Two boats at known locations can also use triangulation to locate and track a third boat or submarine. The distance over which these techniques can be used is limited by temperature gradients in the water, which bend the beam away from the surface and create shadow regions. (See above *Circular and spherical waves: Refraction.*) One of the advantages of ultrasonic waves over sound waves in underwater applications is that, because of their higher frequencies (or shorter wavelengths), the former will travel greater distances with less diffraction.

Uses of sonic ranging

Ranging has also been used to map the bottom of the ocean, providing depth charts that are commonly used in navigation, particularly near coasts and in shallow waterways. Even small boats are now equipped with sonic ranging devices that determine and display the depth of the water so that the navigator can keep the boat from beaching on submerged sandbars or other shallow points. Modern fishing boats use ultrasonic ranging devices to locate schools of fish, substantially increasing their efficiency.

Even in the absence of visible light, bats can guide their flight and even locate flying insects (which they consume in flight) through the use of sonic ranging. Ultrasonic echolocation has also been used in traffic control applications and in counting and sorting items on an assembly line. Ultrasonic ranging provides the basis of the eye and vision systems for robots, and it has a number of important medical applications (see below).

The Doppler effect. If an ultrasonic wave is reflected off a moving obstacle, the frequency of the resulting wave will be changed, or Doppler-shifted. More specifically, if the obstacle is moving toward the source, the frequency of the reflected wave will be increased; and if the obstacle is moving away from the source, the frequency of the reflected wave will be decreased. The amount of the frequency shift can be used to determine the velocity of the moving obstacle. Just as the Doppler shift for radar, an electromagnetic wave, can be used to determine the speed of a moving car, so can the speed of a moving submarine be determined by the Doppler shift of a sonar beam. An important industrial application is the ultrasonic flow meter, in which reflecting ultrasound off a flowing liquid leads to a Doppler shift that is calibrated to provide the flow rate of the liquid. This technique also has been applied to blood flow in arteries. Many burglar alarms, both for home use and for use in commercial buildings, employ the ultrasonic Doppler shift principle. Such alarms cannot be used where pets or moving curtains might activate them.

Materials testing. Nondestructive testing involves the use of ultrasonic echolocation to gather information on the integrity of mechanical structures. Since changes in the material present an impedance mismatch from which an ultrasonic wave is reflected, ultrasonic testing can be used to identify faults, holes, cracks, or corrosion in materials, to inspect welds, to determine the quality of poured concrete, and to monitor metal fatigue. Owing to the mechanism by which sound waves propagate in metals, ultrasound can be used to probe more deeply than any other form of radiation. Ultrasonic procedures are used to perform in-service inspection of structures in nuclear reactors.

Structural flaws in materials can also be studied by subjecting the materials to stress and looking for acoustic emissions as the materials are stressed. Acoustic emission, the general name for this type of nondestructive study, has developed as a distinct field of acoustics.

High-intensity applications. High-intensity ultrasound has achieved a variety of important applications. Perhaps the most ubiquitous is ultrasonic cleaning, in which ultrasonic vibrations are set up in small liquid tanks in which objects are placed for cleaning. Cavitation of the liquid by the ultrasound, as well as the vibration, create turbulence in the liquid and result in the cleaning action. Ultrasonic cleaning is very popular for jewelry and has also been used with such items as dentures, surgical instruments, and small machinery. Degreasing is often enhanced by ultrasonic cleaning. Large-scale ultrasonic cleaners have also been developed for use in assembly lines.

Ultrasonic cleaning

Ultrasonic machining employs the high-intensity vibrations of a transducer to move a machine tool. If necessary, a slurry containing carborundum grit may be used; diamond tools can also be used. A variation of this technique is ultrasonic drilling, which makes use of pneumatic vibrations at ultrasonic frequencies in place of the standard rotary drill bit. Holes of virtually any shape can be drilled in hard or brittle materials such as glass, germanium, or ceramic.

Ultrasonic soldering has become important, especially for soldering unusual or difficult materials and for very clean applications. The ultrasonic vibrations perform the function of cleaning the surface, even removing the oxide layer on aluminum so that the material can be soldered. Because the surfaces can be made extremely clean and free from the normal thin oxide layer, soldering flux becomes unnecessary.

Chemical and electrical uses. The chemical effects of ultrasound arise from an electrical discharge that accompanies the cavitation process. This forms a basis for ultrasound's acting as a catalyst in certain chemical reactions, including oxidation, reduction, hydrolysis, polymerization and depolymerization, and molecular rearrangement. With ultrasound, some chemical processes can be carried out more rapidly, at lower temperatures, or more efficiently.

The ultrasonic delay line is a thin layer of piezoelectric material used to produce a short, precise delay in an electrical signal. The electrical signal creates a mechanical vibration in the piezoelectric crystal that passes through the crystal and is converted back to an electrical signal. A very precise time delay can be achieved by constructing a crystal with the proper thickness. These devices are employed in fast electronic timing circuits.

Medical applications. Although ultrasound competes with other forms of medical imaging, such as X-ray techniques and magnetic resonance imaging, it has certain desirable features—for example, Doppler motion study—that the other techniques cannot provide. In addition, among the various modern techniques for the imaging of internal organs, ultrasonic devices are by far the least expensive. Ultrasound is also used for treating joint pains and for treating certain types of tumours for which it is desirable to produce localized heating. A very effective use of ultrasound deriving from its nature as a mechanical vibration is the elimination of kidney and bladder stones.

Diagnosis. Much medical diagnostic imaging is carried out with X rays. Because of the high photon energies of the X ray, this type of radiation is highly ionizing—that is, X rays are readily capable of destroying molecular bonds in the body tissue through which they pass. This destruction can lead to changes in the function of the tissue involved or, in extreme cases, its annihilation.

One of the important advantages of ultrasound is that it is a mechanical vibration and is therefore a nonionizing form of energy. Thus, it is usable in many sensitive

circumstances where X rays might be damaging. Also, the resolution of X rays is limited owing to their great penetrating ability and the slight differences between soft tissues. Ultrasound, on the other hand, gives good contrast between various types of soft tissue.

Ultrasonic scanning

Ultrasonic scanning in medical diagnosis uses the same principle as sonar. Pulses of high-frequency ultrasound, generally above one megahertz, are created by a piezoelectric transducer and directed into the body. As the ultrasound traverses various internal organs, it encounters changes in acoustic impedance, which cause reflections. The amount and time delay of the various reflections can be analyzed to obtain information regarding the internal organs. In the B-scan mode, a linear array of transducers is used to scan a plane in the body, and the resultant data is displayed on a television screen as a two-dimensional plot. The A-scan technique uses a single transducer to scan along a line in the body, and the echoes are plotted as a function of time. This technique is used for measuring the distances or sizes of internal organs. The M-scan mode is used to record the motion of internal organs, as in the study of heart dysfunction. Greater resolution is obtained in ultrasonic imaging by using higher frequencies—*i.e.*, shorter wavelengths. A limitation of this property of waves is that higher frequencies tend to be much more strongly absorbed.

Because it is nonionizing, ultrasound has become one of the staples of obstetric diagnosis. During the process of drawing amniotic fluid in testing for birth defects, ultrasonic imaging is used to guide the needle and thus avoid damage to the fetus or surrounding tissue. Ultrasonic imaging of the fetus can be used to determine the date of conception, to identify multiple births, and to diagnose abnormalities in the development of the fetus.

Ultrasonic Doppler techniques have become very important in diagnosing problems in blood flow. In one technique, a three-megahertz ultrasonic beam is reflected off typical oncoming arterial blood with a Doppler shift of a few kilohertz—a frequency difference that can be heard directly by a physician. Using this technique, it is possible to monitor the heartbeat of a fetus long before a stethoscope can pick up the sound. Arterial diseases such as arteriosclerosis can also be diagnosed, and the healing of arteries can be monitored following surgery. A combination of B-scan imaging and Doppler imaging, known as duplex scanning, can identify arteries and immediately measure their blood flow; this has been extensively used to diagnose heart valve defects.

Using ultrasound with frequencies up to 2,000 megahertz, which has a wavelength of 0.75 micrometre in soft tissues (as compared with a wavelength of about 0.55 micrometre for light), ultrasonic microscopes have been developed that rival light microscopes in their resolution. The distinct advantage of ultrasonic microscopes lies in their ability to distinguish various parts of a cell by their viscosity. Also, because they require no artificial contrast mediums, which kill the cells, acoustic microscopy can study actual living cells.

Therapy and surgery. Because ultrasound is a mechanical vibration and can be well focused at high frequencies, it can be used to create internal heating of localized tissue without harmful effects on nearby tissue. This technique can be employed to relieve pains in joints, particularly in the back and shoulder. Also, research is now being carried out in the treatment of certain types of cancer by local heating, since focusing intense ultrasonic waves can heat the area of a tumour while not significantly affecting surrounding tissue.

Trackless surgery

Trackless surgery—that is, surgery that does not require an incision or track from the skin to the affected area—has been developed for several conditions. Focused ultrasound has been used for the treatment of Parkinson's disease by creating brain lesions in areas that are inaccessible to traditional surgery. A common application of this technique is the destruction of kidney stones with shock waves formed by bursts of focused ultrasound. In some cases, a device called an ultrasonic lithotripter focuses the ultrasound with the help of X-ray guidance, but a more common technique for destruction of kidney stones, known as endoscopic ultrasonic disintegration, uses a small metal rod inserted through the skin to deliver ultrasound in the 22- to 30-kilohertz frequency region.

INFRASONICS

The term infrasonics refers to waves of a frequency below the range of human hearing—*i.e.*, below about 20 hertz. Such waves occur in nature in earthquakes, waterfalls, ocean waves, volcanoes, and a variety of atmospheric phenomena such as wind, thunder, and weather patterns. Calculating the motion of these waves and predicting the weather using these calculations, among other information, is one of the great challenges for modern high-speed computers.

Aircraft, automobiles, or other rapidly moving objects, as well as air handlers and blowers in buildings, also produce substantial amounts of infrasonic radiation. Studies have shown that many people experience adverse reactions to large intensities of infrasonic frequencies, developing headaches, nausea, blurred vision, and dizziness. On the other hand, a number of animals are sensitive to infrasonic frequencies, as indicated in Table 7. It is believed by many zoologists that this sensitivity in animals such as elephants may be helpful in providing them with early warning of earthquakes and weather disturbances. It has been suggested that the sensitivity of birds to infrasound aids their navigation and even affects their migration.

S-waves, P-waves, and L-waves

One of the most important examples of infrasonic waves in nature is in earthquakes. Three principal types of earthquake wave exist: the S-wave, a transverse body wave; the P-wave, a longitudinal body wave; and the L-wave, which propagates along the boundary of stratified mediums. L-waves, which are of great importance in earthquake engineering, propagate in a similar way to water waves, at low velocities that are dependent on frequency. S-waves are transverse body waves and thus can only be propagated within solid bodies such as rocks. P-waves are longitudinal waves similar to sound waves; they propagate at the speed of sound and have large ranges.

When P-waves propagating from the epicentre of an earthquake reach the surface of the Earth, they are converted into L-waves, which may then damage surface structures. The great range of P-waves makes them useful in identifying earthquakes from observation points a great distance from the epicentre. In many cases, the most severe shock from an earthquake is preceded by smaller shocks, which provide advance warning of the greater shock to come. Underground nuclear explosions also produce P-waves, allowing them to be monitored from any point in the world if they are of sufficient intensity.

The reflection of man-made seismic shocks has helped to identify possible locations of oil and natural-gas sources. Distinctive rock formations in which these minerals are likely to be found can be identified by sonic ranging, primarily at infrasonic frequencies.

ENVIRONMENTAL NOISE

Many forms of noise in the urban environment, including traffic and airplane noise, industrial noise, and noise from electronically amplified music performed at high audio levels in confined rooms, may contribute to hearing damage. Even when the noise level in a working environment may not be dangerous, it can be distracting for those who work in that environment and therefore lead to reduced work production. In addition to the sound level, the character of the noise may be important. Identifiable noises, such as talking or music, may be more distracting for many people than noise produced by air handlers, small motors, or traffic.

Low levels of noise may be overcome using additional absorbing material, such as heavy drapery or sound-absorbent tiles in enclosed rooms. Where low levels of identifiable noise may be distracting, or where privacy of conversations in adjacent offices and reception areas may be important, the undesirable sounds may be masked. A small white-noise source such as static or rushing air, placed in the room, can mask the sounds of conversation from adjacent rooms without being offensive or dangerous to the ears of people working nearby. This type of device

is often used in offices of doctors and other professionals. Another technique for reducing personal noise level is through the use of hearing protectors, which are held over the ears in the same manner as an earmuff. By using commercially available earmuff-type hearing protectors, a decrease in sound level can be attained ranging typically from about 10 decibels at 100 hertz to over 30 decibels for frequencies above 1,000 hertz.

Environmental and industrial noise is regulated in the United States under the Occupational Safety and Health Act of 1970 and the Noise Control Act of 1972. Under these acts, the Occupational Safety and Health Administration has set up industrial noise criteria in order to provide limits on the intensity of sound exposure and on the time duration for which that intensity may be allowed. Maximum daily exposure to noise at various levels is given in Table 8. If an individual is exposed to various levels of noise for different time intervals during the day, the total exposure or dose (D) of noise is obtained from the relation

Calculating total noise exposure

$$D = \left(\frac{C_1}{T_1}\right) + \left(\frac{C_2}{T_2}\right) + \left(\frac{C_3}{T_3}\right) + \ldots, \qquad (41)$$

where C is the actual time of exposure and T is the allowable time of exposure at any level in Table 8. Using this formula, the maximum allowable daily noise dose will be 1, and any daily exposure over 1 is unacceptable.

Criteria for indoor noise are summarized in three sets of specifications that have been derived by collecting subjective judgments from a large sampling of people in a variety of specific situations. These have developed into the noise criteria (NC) and preferred noise criteria (PNC) curves, which provide limits on the level of noise introduced into the environment. The NC curves, developed in 1957, aim to provide a comfortable working or living environment by specifying the maximum allowable level of noise in octave bands over the entire audio spectrum. The complete set of 11 curves specifies noise criteria for a broad range of situations. The PNC curves, developed in 1971, add limits on low-frequency rumble and high-frequency hiss; hence, they are preferred over the older NC standard. Summarized in the curves shown in Figure 13, these criteria provide design goals for noise levels for a variety of different purposes. Part of the specification of a work or living environment is the appropriate PNC curve; in the event that the sound level exceeds PNC limits, sound-absorptive materials can be introduced into the environment as necessary to meet the appropriate standards.

Table 8: Daily Maximum Noise Exposure Permitted by the U.S. Occupational Safety and Health Act of 1970

sound level (decibels)	maximum hours per day
115	< 1/4
110	1/2
105	1
100	2
97	3
95	4
92	6
90	8

Outdoor noise limits are also important for human comfort. Standard house construction will provide some shielding from external sounds if the house meets minimum standards of construction and if the outside noise level falls within acceptable limits. These limits are generally specified for particular periods of the day—for example, during daylight hours, during evening hours, and at night during sleeping hours. Because of refraction in the atmosphere owing to the nighttime temperature inversion, relatively loud sounds can be introduced into an area from a rather distant highway, airport, or railroad. One interesting technique for control of highway noise is the erection of noise barriers alongside the highway, separating the highway from adjacent residential areas. The effectiveness of such barriers is limited by the diffraction of sound, which is greater at the lower frequencies that often predominate in road noise, especially from large vehicles. In order to be effective, they must be as close as possible to either the source or the observer of the noise (preferably to the source), thus maximizing the diffraction that would be necessary for the sound to reach the observer. Another requirement for this type of barrier is that it must also limit the amount of transmitted sound in order to bring about significant noise reduction.

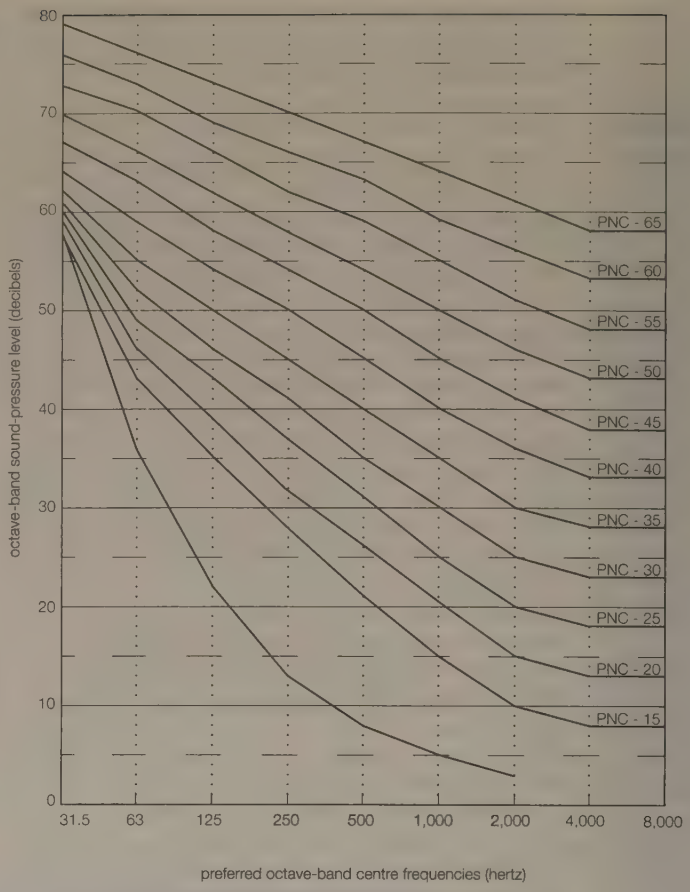


Figure 13: *Preferred noise criteria (PNC) curves.* The lowest curve indicates the approximate threshold of hearing for continuous noise.

From J.E.K. Foreman, *Sound Analysis and Noise Control* (1990), Van Nostrand Reinhold

BIBLIOGRAPHY

General works. Comprehensive discussions of the propagation and perception of sound, containing sections on the ear, on sound recording and reproduction, and on architectural acoustics, are offered in the following books, which require almost no mathematical background: JOHN BACKUS, *The Acoustical Foundations of Music,* 2nd ed. (1977); MURRAY CAMPBELL and CLIVE GREATED, *The Musician's Guide to Acoustics* (1987); JOHN R. PIERCE, *The Science of Musical Sound,* rev. ed. (1992); MICHAEL J. MORAVCSIK, *Musical Sound: An Introduction to the Physics of Music* (1987); and IAN JOHNSTON, *Measured Tones: The Interplay of Physics and Music* (1989). Books requiring an elementary understanding of mathematics include HARVEY E. WHITE and DONALD H. WHITE, *Physics and Music: The Science of Musical Sound* (1980); RICHARD E. BERG and DAVID G. STORK, *The Physics of Sound* (1982), with separate sections demanding considerable knowledge of musical notation and instruments; WILLIAM J. STRONG and GEORGE R. PLITNIK, *Music, Speech, High-Fidelity,* 2nd ed. (1983); JOHN S. RIGDEN, *Physics and the Sound of Music,* 2nd ed. (1985); and DONALD E. HALL, *Musical Acoustics,* 2nd ed. (1991). A somewhat higher level of mathematics is needed for the comprehensive ARTHUR H. BENADE, *Fundamentals of Musical Acoustics* (1976, reissued 1990), a relatively sophisticated classic in the field; and THOMAS D. ROSSING, *The Science of Sound,* 2nd ed. (1990), covering virtually every area of acoustics.

Important advanced texts include the following classics: LEO L. BERANEK, *Acoustics* (1954, reissued 1986); R. BRUCE LINDSAY, *Mechanical Radiation* (1960); and HARRY F. OLSON, *Music, Physics, and Engineering,* 2nd ed. (1967). More recent advanced comprehensive studies are ALLAN D. PIERCE, *Acoustics: An Introduction to Its Physical Principles and Applications* (1981, reissued 1989); F.B. STUMPF, *Analytical Acoustics* (1980); LAWRENCE E. KINSLER *et al., Fundamentals of Acoustics,* 3rd ed. (1982); DONALD E. HALL, *Basic Acoustics* (1987); and S.N. SEN,

Acoustics, Waves and Oscillations (1990). THOMAS D. ROSSING (ed.), *Musical Acoustics* (1988); and CARLEEN MALEY HUTCHINS (ed.), *The Physics of Music: Readings from Scientific American* (1978), are collections of articles.

An enormous amount of physical data on such topics as the velocity of sound and the elastic properties of materials, as well as surveys of important theories in the field, are found in the following reference books: HERBERT L. ANDERSON (ed.), *A Physicist's Desk Reference* (1989); DWIGHT E. GRAY (ed.), *American Institute of Physics Handbook*, 3rd ed. (1972); and RITA G. LERNER and GEORGE L. TRIGG (eds.), *Encyclopedia of Physics*, 2nd ed. (1991). For biographies of scientists who worked in the field of acoustics, see CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography*, 16 vol. (1970–80).

Contemporary research in areas related to sound and its application is covered in periodicals: see *The Journal of the Acoustical Society of America* (monthly); *Acustica* (monthly); *Journal of Sound and Vibration* (biweekly); *Ultrasonics* (quarterly); *Soviet Physics: Acoustics* (bimonthly); and periodicals from other fields of knowledge, such as *JAMA: Journal of the American Medical Association* (weekly), for medical applications of ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* (bimonthly) covers topics of ongoing scientific conferences in all areas of sound and ultrasonics.

History of acoustics. JOHN WILLIAM STRUTT (BARON RAYLEIGH), *The Theory of Sound*, 2nd ed., rev. and enlarged, 2 vol. (1894–96, reissued 1945), remains a most important historical authority on nearly all aspects of theoretical acoustics. HERMAN L.F. HELMHOLTZ, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 2nd English ed. (1885, reprinted 1954; originally published in German, 4th German ed., 1877), is the historical magnum opus in the field of psychoacoustics. Excellent collections of papers of historical interest include R. BRUCE LINDSAY (ed.), *Acoustics: Historical and Philosophical Development* (1973), and *Physical Acoustics* (1974); and STEPHEN G. BRUSH (ed.), *History of Physics: Selected Reprints* (1988).

The ear and hearing. A most important modern work on the physiology of hearing is GEORG VON BÉKÉSY, *Experiments in Hearing* (1960, reprinted 1980). JUAN G. ROEDERER, *Introduction to the Physics and Psychophysics of Music*, 2nd ed. (1975), thoroughly and clearly discusses the ear and hearing, using only basic mathematics. An excellent survey of psychoacoustics is provided in BRIAN C.J. MOORE, *An Introduction to the Psychology of Hearing*, 3rd ed. (1989). Modern experiments in hearing are described in REINIER PLOMP, *Aspects of Tone Sensation: A Psychophysical Study* (1976).

Sound in animals and birds. Data on hearing ranges in animals is collected in RICHARD R. FAY, *Hearing in Vertebrates: A Psychophysics Databook* (1988). CHANDLER S. ROBBINS, BERTEL BRUUN, and HERBERT S. ZIM, *Birds of North America*, expanded rev. ed. (1983), includes audio spectrographs of bird calls.

Electromechanical transducers, sound recording, and sound reproduction. HARRY F. OLSON, *Acoustical Engineering* (1957), and *Modern Sound Reproduction* (1972), are the advanced classics in the field, including detailed discussions of loudspeaker design. An excellent introduction to audio equipment is provided in INSTITUTE OF HIGH FIDELITY, *Official Guide to High Fidelity*, 2nd ed. (1978). A later introduction to audio reproduction is offered in KENNETH W. JOHNSON, WILLARD C. WALKER, and JOHN D. CUTNELL, *The Science of Hi-Fidelity*, 2nd ed. (1981). Advanced treatment of electromechanical transducers is found in JOSEF MERHAUT, *Theory of Electroacoustics*, trans. from Czech (1981); JOHN BORWICK, *Microphones: Technology and Technique* (1990); and MARTIN COLLOMS, *High Performance Loudspeakers*, 4th ed. (1991). JOHN EARGLE, *Sound Recording*, 2nd ed. (1980), is of interest for its treatment of older audio reproduction technology, including LP disc recordings and quadraphonic sound. Eargle's later work, *Handbook of Recording Engineering*, 2nd ed. (1992), also covers digital sound recording. Both audio and video magnetic tape technology is discussed in JOHN C. MALLINSON, *The Foundations of Magnetic Recording* (1987); and C. DENIS MEE and ERIC D. DANIEL, *Magnetic Recording*, 3 vol. (1987–88). For digital techniques, see KEN C. POHLMANN, *Principles of Digital Audio*, 2nd ed. (1989), on the compact disc; and JOHN WATKINSON, *RDAT* (1991), on the rotary head digital audiotape.

Architectural acoustics. An important survey of the subject is found in WALLACE C. SABINE, *Collected Papers on Acoustics* (1964). LEO L. BERANEK (ed.), *Noise and Vibration Control*, rev. ed. (1988), contains excellent sections applying to concert halls; and LEO L. BERANEK, *Music, Acoustics & Architecture* (1962, reprinted 1979), discusses more than 50 existing concert halls, relating physical properties of sound waves in auditoriums to their subjective effects. Eighty-seven concert halls are surveyed in the illustrated work by RICHARD H. TALASKE, EWART A. WETHERILL, and WILLIAM J. CAVANAUGH (eds.), *Halls for Music Performance: Two Decades of Experience, 1962–1982* (1982). LOTHAR CREMER and HELMUT A. MÜLLER, *Principles and Applications of Room Acoustics*, 2 vol. (1982; originally published in German, 2nd ed., 1976–78), is a detailed and advanced treatment.

Ultrasonics. Classic works in the field of ultrasonics include BASIL BROWN and JOHN E. GOODMAN, *High-Intensity Ultrasonics: Industrial Applications* (1965); ROBERT T. BEYER and STEPHEN V. LETCHER, *Physical Ultrasonics* (1969); DALE ENSMINGER, *Ultrasonics: Fundamentals, Technology, Applications*, 2nd ed., rev. and expanded (1988); and ROBERT T. BEYER, *Nonlinear Acoustics* (1974). P.N.T. WELLS, *Biomedical Ultrasonics* (1977), provides a summary of biomedical applications through the time of publication. Another survey of developments of the period is ROBERT T. BEYER, "A New Wave of Acoustics," *Physics Today*, 34(11):145–157 (November 1981). JAMES R. MATTHEWS (ed.) *Acoustic Emission* (1983), describes in great detail modern techniques for testing materials with ultrasonic emissions. A variety of applications are studied in A.P. CRACKNELL, *Ultrasonics* (1980); KENNETH S. SUSLICK (ed.), *Ultrasound: Its Chemical, Physical, and Biological Effects* (1988); and D. STANSFIELD, *Underwater Electroacoustic Transducers: A Handbook for Users and Designers* (1990). B.P. HILDEBRAND and B.B. BRENDEN, *An Introduction to Acoustical Holography* (1972), surveys holographic techniques and the basis for later development in medical imaging. HARVEY FEIGENBAUM, *Echocardiography*, 4th ed. (1986), discusses ultrasonic cardiography. RUSSEL K. HOBBIE (ed.), *Medical Physics: Selected Reprints* (1986), collects articles on advances in medical ultrasonics. Information on later research activity in the field is found in B.R. MCAVOY (ed.), *IEEE 1990 Ultrasonics Symposium: Proceedings*, 3 vol. (1989); and in the materials published in the serial *Physical Acoustics: Principles and Methods* (irregular), ed. by WARREN P. MASON and R.N. THURSTON.

Infrasonics. Infrasonic waves in nature are studied in EARL E. GOSSARD and WILLIAM H. HOOKE, *Waves in the Atmosphere: Atmospheric Infrasound and Gravity Waves: Their Generation and Propagation* (1975). Experimental infrasonic techniques are discussed in A.F. YAKUSHOVA, *Geology with the Elements of Geomorphology* (1986; originally published in Russian, 2nd rev. ed., 1983); and BRUCE A. BOLT (ed.), *Earthquakes and Volcanoes: Readings from Scientific American* (1980).

Environmental noise. U.S. law and public policy regarding environmental noise are described in UNITED STATES OFFICE OF NOISE ABATEMENT AND CONTROL, *Public Health and Welfare Criteria for Noise* (1973). A wealth of data is accumulated in CYRIL M. HARRIS (ed.), *Handbook of Noise Control*, 2nd ed. (1979). Noise control applications are examined in P.O.A.L. DAVIES, M. HECKL, and G.L. KOOPMAN, *Noise Generation and Control in Mechanical Engineering* (1982); LEWIS H. BELL *et al.*, *Industrial Noise Control: Fundamentals and Applications* (1982); and JOHN E.K. FOREMAN, *Sound Analysis and Noise Control* (1990). (R.E.B.)

South America

South America, the fourth largest of the world's continents, is the southern portion of the landmass generally referred to as the New World, the Western Hemisphere, or simply the Americas. It is compact and roughly triangular in shape, being broad in the north and tapering to a point—Cape Horn, Chile—in the south.

South America is bounded by the Caribbean Sea to the northwest and north, the Atlantic Ocean to the northeast, east, and southeast, and the Pacific Ocean to the west. In the northwest it is joined to North America by the Isthmus of Panama, which forms a land bridge narrowing to about 50 miles (80 kilometres) at one point. Drake Passage, south of Cape Horn, separates South America from Antarctica.

Relatively few islands are associated with the continent, except in the south. These include the glaciated coastal archipelagoes of Argentina and Chile. The Falkland (Malvinas) Islands are east of southern Argentina. To the north, the West Indies stretch from Trinidad to Florida, but these islands usually are associated with North America. Of the remainder, most are small oceanic islands off the coasts of South America, including the Galápagos Islands, Ecuador, in the Pacific Ocean.

South America has a total area of about 6,878,000 square miles (17,814,000 square kilometres), or roughly one-eighth of the land surface of the Earth. Its greatest north-south extent is about 4,700 miles, from Point Gallinas, Colom., to Cape Horn; while its greatest east-west extent is some 3,300 miles, from Cape Branco, Braz., to Point Pariñas, Peru. At 22,831 feet (6,959 metres) above sea level, Mount Aconcagua, in Argentina, near the border with Chile, is not only the continent's highest point but also the highest elevation in the Western Hemisphere. The Valdés Peninsula, on the southeastern coast of Argentina, includes the lowest point, at 131 feet (40 metres) below sea level. In relation to its area, the continent's coastline—some 15,800 miles in length—is exceptionally short.

The name America is derived from that of the Italian merchant and navigator Amerigo Vespucci, one of the earliest European explorers of the New World. The term America originally was applied only to South America, but the designation soon was applied to the entire landmass. Because Mexico and Central America share an Iberian heritage with nearly all of South America, this entire region frequently is grouped under the name Latin America.

South America's geologic structure consists of two dissymetric parts. In the larger, eastern portion are found a number of stable shields forming highland regions, separated by large basins (including the vast Amazon basin). The western portion is occupied almost entirely by the Andes Mountains. The Andes—formed as the South American Plate moved westward and forced the oceanic plate to the west under it—constitute a gigantic backbone along the entire Pacific coast of the continent. The basins east of the Andes and between the eastern highlands have been filled with immense quantities of sediment washed down by the continent's great rivers and their tributaries.

No other continent—except Antarctica—penetrates so far to the south. Although the northern part of South America extends north of the Equator and four-fifths of its landmass is located within the tropics, it also reaches subantarctic latitudes. Much of the high Andes lie within the tropics but include extensive zones of temperate or cold climate in the vicinity of the Equator—a circumstance that is unique. The great range in elevation produces an unrivaled diversity of climatic and ecological zones, which is probably the most prominent characteristic of South American geography.

The original inhabitants of South America are believed to have descended from the same Asiatic peoples who migrated to North America from Siberia during the most recent (Wisconsin) ice age. Few of these peoples, however, survived the arrival of Europeans after 1500, most succumbing to disease or mixing with people of European and (especially in Brazil) African origin. Some parts of the continent are now industrialized, with modern cities, but the people of most regions still follow an agricultural way of life. The wealth of mineral products and renewable resources is considerable, yet economic development in most of the continent lags behind the more industrially advanced regions of the world. Nonetheless, considerable concern has arisen about the rapidly increasing and often destructive exploitation of these resources.

(J.P.D./C.W.M.)

This article treats the physical and human geography of South America, followed by discussion of geographic features of special interest. For discussion of individual countries of the continent, see specific articles by name, *e.g.*, ARGENTINA, BRAZIL, and VENEZUELA. For discussion of major cities of the continent, see the articles BUENOS AIRES, CARACAS, LIMA, RIO DE JANEIRO, and SÃO PAULO. For discussion of the indigenous peoples of the continent, see the articles AMERICAN PEOPLES, NATIVE; and PRE-COLUMBIAN CIVILIZATIONS. Related topics are discussed in the articles AMERICAN PEOPLES, ARTS OF NATIVE; LATIN AMERICA, THE HISTORY OF; and LATIN-AMERICAN LITERATURE. For further references, see the *Index*.

The article is divided into the following sections:

Physical and human geography 579
- Geologic history 579
  - The Precambrian 579
    - The Trans-Amazonian cycle
    - The Brazilian cycle
  - The Paleozoic Era 582
    - Early Paleozoic events
    - The formation of Pangaea
  - The Mesozoic and Cenozoic eras 583
    - Events in the Mesozoic
    - The Andean orogeny
    - Present geologic setting
- The land 585
  - Relief 585
    - The Andes Mountains
    - The plateaus
    - The lowlands
  - Drainage 587
    - Rivers
    - Lakes
    - Marshes and swamps
  - Soils 588
  - Climate 589
    - Factors influencing climate
    - Climatic regions
  - Plant life 591
    - Vegetation zones
    - Human influences on vegetation
  - Animal life 593
    - Principal faunal types
    - Ecological communities
    - Human influences on wildlife
- The people 595
  - Ethnic origins and migrations 595
    - Indians
    - Iberians
    - Africans
    - Postindependence overseas immigrants
  - Population and ecological distribution 596
    - The present population
    - Culture areas
    - Linguistic patterns
    - Religious patterns
    - Sociological changes

Demographic patterns 597
The demographic transition and fertility
Effects of rapid population increase
The economy 599
Resources 599
Mineral resources
Biological resources
Forestry, fishing, and mining 602
Agriculture 602
Principal crops
Livestock
Industry 604
Power and irrigation 604
Trade 604
Internal trade
External trade
Transportation 605
Roads
Railways
Maritime transport
Waterways
Air transportation
South American geographic features of special interest 606
Landforms 606
Andes Mountains 606
Gran Chaco 611
Patagonia 614
Drainage systems 615
Amazon River basin 615
Orinoco River basin 622
Río de la Plata system 625
São Francisco River 630
Bibliography 632

PHYSICAL AND HUMAN GEOGRAPHY

Geologic history

Three stages of geologic history

The geologic history of South America can be summarized in three different developmental stages, each corresponding to a major division of geologic time. The first stage encompassed Precambrian time (3.96 billion to 570 million years ago) and was characterized by a complex series of amalgamations and dispersals of stable blocks of protocontinental crust called cratons. The second stage coincides with the Paleozoic Era (570 to 245 million years ago), during which time the cratons and material accreted to them contributed to the formation first of the supercontinent Gondwana (or Gondwanaland) and then of the even larger Pangaea. The third stage, in which the present continental structure emerged, occurred in the Mesozoic and Cenozoic eras (the last 245 million years) and includes the breakup of Pangaea and Gondwana, the opening of the South Atlantic Ocean, and the generation of the Andean cordillera.

The present tectonic framework of South America consists of three fundamental units: the ancient cratons, the relatively recent Andean ranges, and a number of basins. Five cratons—Amazonia, São Francisco, Luis Alves, Alto Paraguay, and Río de la Plata—represent the Precambrian core of South America, and (with the exception of the Alto Paraguay craton) these now appear as upwarped massifs arrayed from north to south in the immense eastern portion of the continent; a number of other Precambrian crustal blocks also were accreted along the margins of South America over geologic time. The lofty ranges and intermontane plateaus of the Andes rise along the entire western margin of the continent and represent the collision in the Cenozoic Era (the last 66.4 million years) of the Pacific and South American plates brought about by the opening of the South Atlantic. Finally, vast, downwarped, sediment-filled basins are found between the cratons and along the entire eastern margin of the Andes.

THE PRECAMBRIAN

Precambrian rocks constitute the oldest rocks of the continent and are preserved in the five core cratons. These rocks are represented by high- to low-grade metamorphosed assemblages along heavily deformed belts of plutonic (intrusive), metavolcanic (metamorphosed extrusive igneous rocks), and metasedimentary rocks. Rocks of Archean age (2.5 to 3.8 billion years old) are known in the Amazonia, Luis Alves, and São Francisco cratons, although precisely dated rock samples are scarce. Ages older than 3 billion years have been reported in the Imataca Complex of Venezuela and in the Xingu area of Brazil, both in the Amazonia craton. The oldest rocks found so far—with ages of some 3.4 billion years—are in the São Francisco craton in the Brazilian state of Bahia. In the other cratons (*e.g.*, the Río de la Plata craton in Uruguay) the dating of Archean rocks has been inconclusive. Greenstone belts, which are remnants of Archean oceanic crust emplaced in the suture zones (convergent plate boundaries), contain most of South America's known large gold deposits, such as those located near Belo Horizonte, Braz. Two major cycles of crustal deformation occurred in the Precambrian, widely separated in time from each other. The first, called the Trans-Amazonian cycle, took place approximately 2.2 to 1.8 billion years ago; and the second, the Brazilian cycle, between about 900 and 570 million years ago.

Oldest rocks on the continent

The Trans-Amazonian cycle. Trans-Amazonian rocks can be subdivided into three distinct groups: orogenic belts, such as the Maroni-Itacaiúnas belt of the Amazonia craton or the Salvador-Juazeiro belt of the São Francisco; stable cover rocks, such as the Chapada Diamantina formation in Bahia or the Carajás and Roraima platform deposits; and large extensional dike swarms (groups of tabular intrusions of igneous rock into sedimentary strata). The orogenic belts represent old mountain chains that had been formed either along the margins of the continent as geosynclines (downwarps of the Earth's crust) and then uplifted, such as the Maroni-Itacaiúnas belt, or were the result of collisions between continental blocks, such as the Tandil belt in Buenos Aires, Arg.

Such collisions are believed to have formed a supercontinent (sometimes called the first Pangaea) some 1.8 billion years ago. The sedimentary cover of this supercontinent (preserved on the Amazonia craton), consisting of post-collision rhyolites and clastic shelf deposits, was deep and widespread and obliterated earlier suture boundaries. Extensive stratified iron and manganese deposits are found in these sequences, such as near Carajás, Braz. Early phases of continental-plate dispersal produced extensive dike swarms of mafic rock, including a zone some 60 miles wide in west-central Uruguay where hundreds of gabbro dikes are now emplaced along a 150-mile stretch.

The Brazilian cycle. Rocks of the Brazilian cycle today are manifested in a series of orogenic belts—developed mainly on previously deformed continental crust—that were formed during the amalgamation of the Precambrian cratons into the first supercontinent in late Proterozoic time (900 to 570 million years ago). Most of present-day South America, encompassing the platforms of Brazil, Guyana, and southern Venezuela, was accreted at this time—together with Africa—to form the western part of the huge southern supercontinent of Gondwana; Precambrian blocks that were not part of Gondwana—notably the Santa Marta massif in Colombia, Mount Arequipa in Peru, and Patagonia in Argentina—were accreted later during Paleozoic times.

The Brasilides

The Brasilides in the southern Brazilian state of Matto Grosso represent the type locality of the Brazilian orogenic cycle. There, important sequences of green schists, platform limestones, and quartzites, as well as red bed molasse formations (associated with granitoids), permit a reconstruction of the collision between the Amazonia craton's passive (*i.e.*, without active volcanoes) margin and the Alto Paraguay craton's active margin (now partially covered by the Paraná basin). The interpreted suture zone between these two cratons corresponds to the Paraguay-Araguaia line, along which mafic and ultramafic rocks are found today.


ATLANTIC OCEAN
Cape São Roque
Cape Branco
Cape Norte
SERRA DO ESPINHAÇO
MESTRE UPLANDS
SERRA DA CANASTRA
Brasília
Tucuruí Dam
SERRA DOS CARAJÁS
B R A Z I L
BRAZILIAN HIGHLANDS
MATO GROSSO PLATEAU
SERRA DO CACHIMBO
PANTANAL
PARECIS MOUNTAINS
LLANOS DE CHIQUITOS
BOLIVIA
Sucre
CORDILLERA ORIENTAL
La Paz
Mount Illimani 6322
CORDILLERA REAL
CORDILLERA OCCIDENTAL
A L T I P L A N O
ATACAMA
Cayenne
FRENCH GUIANA
830
TUMUC-HUMAC MTS
Paramaribo
SURINAME
Juliana Top 1230
Georgetown
GUYANA
ACARAI MTS
GUIANA HIGHLANDS
PAKARAIMA MOUNTAINS
Mount Roraima 2772
SIERRA PARIMA
ORINOCO DELTA
IMATACA MTS.
Guri Dam
Mount Bolívar 802
VENEZUELA
Mount Duida 2396
Neblina Peak 3014
Caracas
L L A N O S
SEGOVIA HIGHLANDS
CORDILLERA DE MÉRIDA
Bolívar Peak 5007
5493
SIERRA NEVADA DE COCUY
SIERRA DE PERIJÁ
Bolívar Peak 5775
SEA
Gulf of
GUAJIRA PENINSULA
Bogotá
COLOMBIA
CORDILLERA ORIENTAL
CORDILLERA CENTRAL
CORDILLERA OCCIDENTAL
Mount Huila 5750
Quito
Chimborazo Volcano 6310
ECUADOR
A N D E S
CORDILLERAS DEL CÓNDOR
SECHURA DESERT
Point Pariñas
CORDILLERA ORIENTAL
CORDILLERA DE VILCABAMBA
CORDILLERA DE CARABAYA
PERU
Ticlio Pass
Mount Huascarán 6768
CORDILLERA BLANCA
CORDILLERA NEGRA
CORDILLERA OCCIDENTAL
Mount Chachani 6075
Lima
CHINCHA ISLANDS
10


PACIFIC OCEAN
ATLANTIC OCEAN
ANDES MOUNTAINS
ATACAMA DESERT
ATACAMA PLATEAU
Mount Ojos del Salado 6893
Mount Bonete 6872
PAMPEAN SIERRAS
SIERRA DE CORDOBA
Mount Champaqui 2790
SIERRA DE SAN LUIS
Mount Mercedario 6770
Mount Aconcagua 6959
Bermejo Pass
Santiago
Mount Tupungato 6570
CHILE
Domuyo Volcano 4709
PAMPAS
GRAN CHACO
CHACO BOREAL
CHACO CENTRAL
PARAGUAY
Asuncion
CHACO AUSTRAL
SERRA DA MANTIQUEIRA
SERRA DO MAR
Sugar Loaf 396
Itaipu Dam
Iguaçu Falls
SERRA DE MISIONES
URUGUAY
GRANITE RANGE
Buenos Aires
Montevideo
Mount Catedral 513
Este Point
Point Sur del Cabo San Antonio
ARGENTINA
PATAGONIA
Mount Tronador 3491
VALDES PENINSULA
Corcovado Volcano 2300
Cape Tres Puntas
Mount Fitzroy 3375
FALKLAND ISLANDS (ISLAS MALVINAS) (Administered by U.K., claimed by Arg.)
Stanley
TIERRA DEL FUEGO
ESTADOS ISLAND
Mount Darwin 2438
Cape Horn
Passage
National capitals
International boundaries
Canals
Intermittent rivers
Dams
Waterfalls and rapids
Glaciers
Swamps and marshes
Salt flats
Sand areas
ELEVATION
©Encyclopædia Britannica Inc

Several other Brazilian belts are known, such as the structurally complex Borborema belt and the Dom Feliciano belt in southern Brazil and Uruguay, which resulted from the collision between the Río de la Plata craton and the Kalahari craton of present-day Africa. The Dom Feliciano belt represents a complex suture zone where rocks typical of a late Proterozoic arc system were trapped between the two cratons; these rocks were then covered by plateaus of rhyolites during the Early Cambrian Period (570 to 540 million years ago). A striking coincidence exists between this suture, which is known as the Brazilian–Pan-African suture, and the inception of the future rift system that opened the Atlantic Ocean. The Pampean Sierras in Argentina are a good example of a Brazilian belt formed by accretion of an island-arc system and several small continental plates.

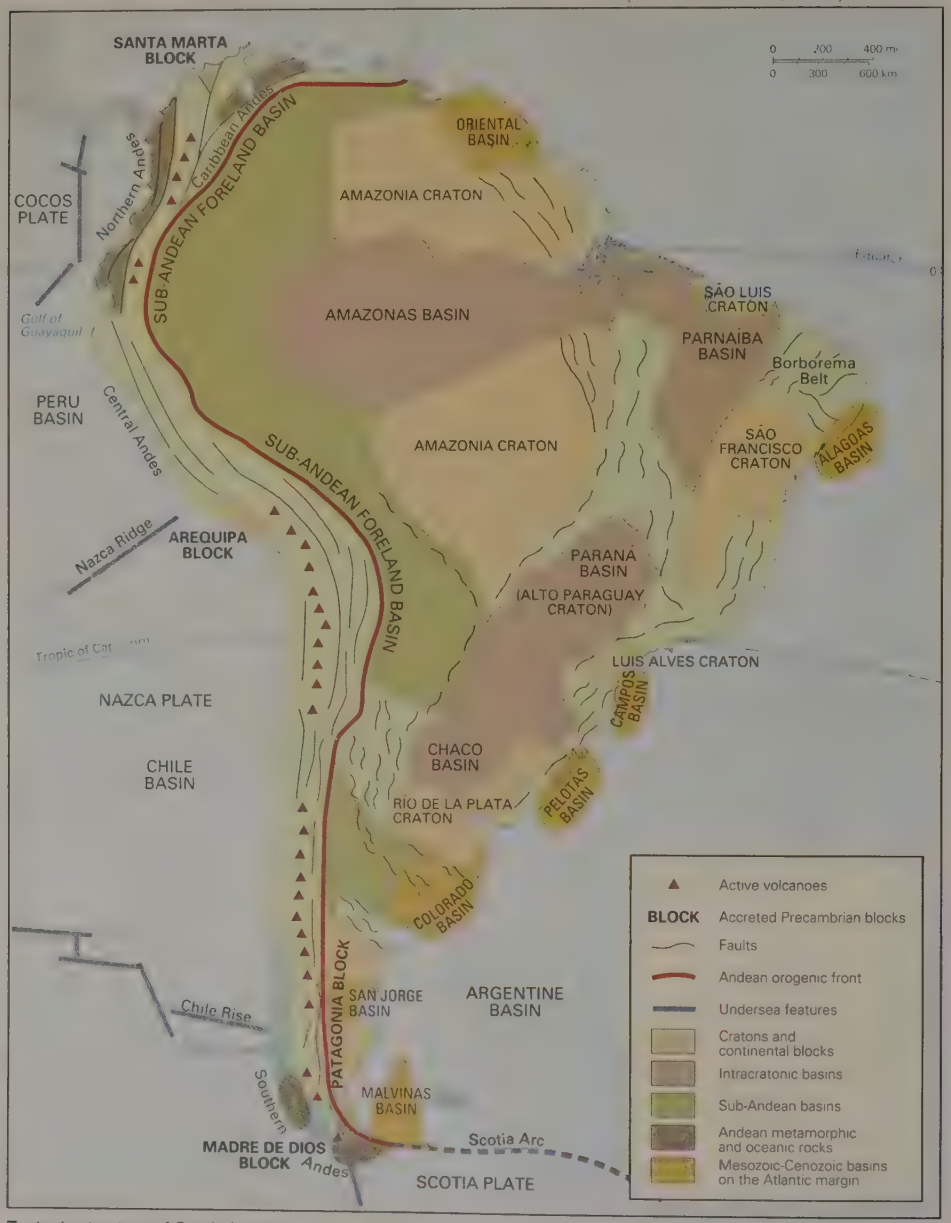
THE PALEOZOIC ERA

Early Paleozoic events. The continent's early Paleozoic rocks depict the breakup of the first supercontinent, an event probably related to the separation of eastern North America from the pre-Andean basement rocks of western South America. As a result of this separation, a series of passive continental margins developed along the western side of the continent from Venezuela and Colombia to central Argentina; essentially, the Precambrian platform amalgamated during the Brazilian cycle. These rifted margins today are represented mainly by clastic rocks from the Late Cambrian and Early Ordovician periods (*i.e.*, roughly 500 million years old) bearing numerous trilobites and graptolites, as in the Cordillera Oriental of Bolivia. The early Paleozoic rift that produced these margins also initiated the development of several large intracratonic basins within the continent (*e.g.*, the Amazonas, Parnaíba, Paraná, and Chaco basins). Thick deposits of sedimentary rocks have since accumulated in these basins.

The passive margins of the early Paleozoic were partially activated by subduction of oceanic crust (*i.e.*, the forced descent of oceanic crust beneath the leading edge of an overriding continental plate) during Late Cambrian–Ordovician times. When the oceanic crust was totally consumed, subduction ceased and a series of small continental blocks collided against the western side of the continent. Allochthonous (transported) continental blocks thus were emplaced in the Cordillera Oriental of Ecuador, Colombia, and Venezuela at the end of the Silurian Period (about 408 million years ago). Rock ages corresponding to those of the North American Grenville orogenic belt (*c.* 1.3 to 1.2 billion years old), as well as affinities to North American fauna of the Devonian Period (408 to 360 mil-

Correspondence to North American rocks

Adapted from Victor A. Ramos, University of Buenos Aires



Tectonic structure of South America.

lion years ago), suggest that these blocks were once part of North America.

Farther south, another series of blocks collided against the continent. These included the Arequipa craton in southern Peru and Bolivia, the Precordillera region of west-central Argentina, and Patagonia in southern Argentina. At the same time, some minor blocks consisting of rocks exhibiting a marine affinity were accreted to the continent in the southern Patagonian archipelago of Chile.

In the course of the subduction process that preceded these collisions, a series of north–south-trending belts of plutonic and volcanic rock formed offshore of the continent and parallel to the coast. Because of the later accretions of continental crust to the coastal margin, these belts were shifted more than 250 miles westward and today form prominent outcrops in northern Patagonia, the western Pampean Sierras, the Cordillera Oriental of Bolivia and northern Argentina, and the Cordillera Oriental of Colombia and Venezuela.

The collision of these blocks also produced a series of peripheral foreland basins, which were the result of crustal deformation and the stacking of slices of basement rocks in the orogenic areas. Examples of basins of the early Paleozoic age are the Beni basin in Bolivia and the Alhuampa and Las Breñas basins in northern Argentina. The late Paleozoic Claromecó foreland basin in northern Patagonia is now occupied by a sedimentary accumulation more than five miles thick that was formed at the same time as the Karoo basin in southern Africa, both basins resulting from the collision of the microcontinent of Patagonia against Gondwana.

The formation of Pangaea. The Paleozoic ended with the final amalgamation of Gondwana, which together with Laurasia to the north constituted the late Paleozoic supercontinent of Pangaea. Subduction beneath the western margin of Pangaea slowly ceased. The igneous rocks formed in the volcanic arc that developed along what is now the Cordillera Central between Chile and Argentina and then along the western continental margin, are transitional in nature—*i.e.,* the composition of the rocks changes from primarily andesitic to predominantly rhyolitic. These changes in mineral composition indicate the passage from a subduction-related compressive regime to one of extensive magmatic activity and crustal extension. Vast sheets of magma, consisting of flood basalts and of rhyolite deposits up to 2.5 miles thick, covered the west from southern Peru to the border between Argentina and Chile. Farther north this activity was partially obliterated in Cenozoic times by the uplifting of the Andes and by volcanic cover.

THE MESOZOIC AND CENOZOIC ERAS

Events in the Mesozoic. The mosaic of continental blocks accreted to form the Pangaea supercontinent was unstable and remained amalgamated for only a few million years. Extensive sedimentary cover indicative of arid conditions was accumulated unevenly in the late Paleozoic basins. Desertic sandstones, mudstones, and tuffs of Triassic age (245 to 208 million years ago) have preserved fossils of a rich fauna of dinosaurs and mammallike reptiles, as in the Ischigualasto basin of Argentina.

A series of Middle to Late Triassic basins also developed through horizontal crustal extension during the early phases of Pangaea's dispersal. These rifted basins largely followed the previous Paleozoic sutures along the western side of the continent. Crustal extension reactivated the inner part of the supercontinent as well, with an increase in subsidence in the Parnaiba and Paraná intracratonic basins, where deposits of Triassic age have been recovered from core samples.

Opening of the South Atlantic

The opening of the South Atlantic Ocean is recorded in a series of Mesozoic and Cenozoic basins that developed along the present Atlantic margin. Most of these basins have clastic red beds that date to the Late Jurassic and Early Cretaceous periods (about 163 to 97.5 million years ago). North of Pôrto Alegre, Braz., are found evaporite salt deposits created in marine basins with restricted circulation. Sediments formed under less restricted drift conditions began approximately 125 million years ago and are younger to the north, where they are mostly represented by clastic marine deposits. The basins of northern Brazil have carbonate deposits mixed with clastics. Deposits laid down in a restricted-circulation, anoxic (oxygen-poor) environment along the basins of Brazil and Argentina now contain abundant black shales rich in organic matter and are an important source of hydrocarbons.

Open marine conditions have prevailed in the Atlantic basins since mid-Cretaceous times. The first open oceanic circulation between the South and North Atlantic oceans was established along the passive margin of South America in the Late Cretaceous Period (97.5 to 66.4 million years ago), though marine sediments had accumulated and been lithified in these basins for some time prior to that.

The Andean orogeny. Coincident with most of the Cenozoic Era has been the Andean orogeny, the most significant geologic event of the era. The mountain ranges, however, display some of the same features found in the previous orogenies that developed along the western continental margin, such as the classical Andean volcanic belt, the east-vergence sub-Andean thrust and fold belt, and a series of cordilleras trending parallel to the Pacific oceanic trench. These features are a response to subduction of the ocean crust that was accelerated by the opening of the South Atlantic; and this subduction overshadows all other geomorphic processes along South America's Pacific margin.

The Andean orogeny has three distinct segments, each of which developed in a different geologic setting. The segments are differentiated by their relative abundances of Mesozoic-Cenozoic, metamorphic, and oceanic rocks and are divided into Northern, Central and Southern sectors.

The Northern Andes. North of the Gulf of Guayaquil in Ecuador and Colombia, a series of accreted oceanic terranes (discrete allochthonous fragments) have developed that constitute the Baudo, or Coastal, Mountains and the Cordillera Occidental. They were accreted during Cretaceous and early Cenozoic times. Structurally composed of oceanic volcanic arcs that were amalgamated after each collision by high-angle, west-verging thrusts, the Northern Andes are characterized by the heavily deformed metamorphic rocks and ophiolitic suites that developed during these collisional episodes. During mid-Cenozoic times, a continental magmatic arc was formed between the eastern and western cordilleras.

Farther east, the Andes of Venezuela (the Caribbean Andes) resulted from the collision of the Caribbean and South American plates during Cretaceous times. This complex setting developed a series of wrench faults and related basins east of Bucaramanga (Colombia) and north of the Orinoco River delta (Venezuela). One of these basins, now occupied by Lake Maracaibo, has the largest accumulation of hydrocarbon deposits so far discovered in South America.

Lake Maracaibo

The Central Andes. The Central Andes lie between the Gulfs of Guayaquil and Penas and thus encompass southern Ecuador, Peru, western Bolivia, and northern and central Argentina and Chile. They are characterized by their continental basement rocks and by an absence of oceanic and metamorphic rocks. The formation of the Central Andes was determined by subduction processes that occurred in the absence of major plate collisions. A period of crustal extension prevailed from the Jurassic Period (208 to 144 million years ago) to Early Cretaceous times, when important volcanic piles and plutonic rocks were emplaced. Back-arc basins developed in the sub-Andean regions, controlled by extensional faulting that occurred at about the same time the South Atlantic was opening.

The middle of the Cretaceous in the Central Andes was marked by a change in tectonic activity—from crustal extension to crustal compression. This change was related to an increase in the convergence rate between South America and the adjacent oceanic plate, which initiated the formation of a series of sub-Andean foreland basins from Colombia to central Argentina. Within these basins are now concentrated most of the petroleum resources of the Andean countries.

Since Cretaceous times the Central Andes have been characterized by considerable volcanism along the axis of

Geologic structure of South America.

the principal cordillera. Andesites, basalts, and rhyolites have been the major rock types to result from this activity, with some granitoids as well. Most of the gold and copper mined in Peru, Bolivia, and Chile comes from these formations.

The Southern Andes. The cordilleras south of the Gulf of Penas constitute the Southern Andes. These belts are defined by a long linear batholith (large exposed mass of coarse-grained igneous rock) that now extends unbroken to Estados Island in the South Atlantic. Outcrops of Early Cretaceous mafic and ultramafic rock found south of latitude 50° S along the axis of the cordillera have been interpreted as ocean floor of a back-arc marginal basin. Metamorphic rocks of Andean age are preserved only in the Darwin Cordillera along the Fuegian Andes of Chile. The eastern sub-Andean belt is composed of a series of back-arc and foreland basins, in which sediments more than five miles thick have accumulated.

Present geologic setting. The glaciations that encompass most of the Pleistocene Epoch (1,600,000 to 10,000

years ago) began in southern South America as early as the Late Miocene Epoch (*i.e.*, about 9,000,000 years ago), when ice caps first covered the Patagonian Andes. Maximum ice expansion was reached about 1,000,000 years ago during the Early Pleistocene, when ice sheets covered the Andes from Ecuador to Tierra del Fuego. In some areas, notably Patagonia, ice extended east to the Atlantic Ocean. About 10,000 years ago the glacial ice retreated, and the present landscape of South America began to take shape. South America's contemporary geology is characterized by continued volcanic and seismic activity along the Andes and relatively aseismic conditions to the east. (V.A.R.)

The land

RELIEF

South America has two major mountain systems of contrasting nature. Bordering the Pacific Ocean to the west, the geologically young cordilleras of the Andes extend along the entire continent from north to south. Stretching along the continent's northern and eastern sides are the ancient Guiana and Brazilian highlands, which are much lower in elevation and slope gently to the west; farther south are the Patagonian plateaus. Lowlands—the basins of the Orinoco, Amazon, and Paraguay-Paraná rivers and the plains of the Pampas—divide the highlands from one another. Taken as a whole, the relief of the continent shows a great imbalance: the major drainage divide is far to the west along the crest of the Andes. Thus, rain that falls only 100 miles (160 kilometres) east of the Pacific may flow to the Atlantic, 2,500 miles away.

The Andes Mountains. The ranges of the Andes Mountains, about 5,500 miles long and second only to the Himalayas in average elevation, constitute a formidable and continuous barrier, with many summits exceeding 20,000 feet (6,100 metres). The Venezuelan Andes—the northernmost range of the system—run parallel to the Caribbean coast west of Caracas, before turning to the southwest and entering Colombia. In Colombia the Andes—now trending generally to the north and south—form three distinct massifs: the Cordilleras Oriental, Central, and Occidental. The valley of the Magdalena River, between the Oriental and the Central ranges, and the valley of the Cauca River, between the Central and the Occidental ranges, are huge rift valleys formed by faulting rather than by erosion. An aerial view of the Andes in Colombia shows, within relatively short distances, a succession of hot lowlands interspersed with high ranges with snowcapped peaks.

The Northern Andes

In Ecuador the Andes form two parallel cordilleras, one facing the Pacific and the other descending abruptly eastward toward the Amazon basin, crowned by towering peaks. Between the ranges lies a series of high basins. These ranges continue southward into Peru, where the relief becomes considerably more complex; the highest of the Peruvian peaks is Mount Huascarán (22,205 feet) in the Cordillera Blanca.

Beginning to the south of Lima and extending through western Bolivia, the Andes again form two distinct ranges. Between them lies the Altiplano, a vast complex of high plateaus between 12,000 and 15,000 feet in elevation and as much as 125 miles wide. The Altiplano forms a maze of valleys, hills, and vast plains without equivalent except in Tibet. Water accumulates in closed basins to form marshes and lakes, the largest of which is Lake Titicaca on the border of Peru and Bolivia. This central region of the Andes has been dissected by several large rivers, all of which have cut spectacular gorges down the eastern slopes.

Farther to the south—along the border between Chile and Argentina—the Andes form a single but complex chain with many of the system's highest peaks, including Mount Aconcagua; south of Aconcagua, elevations gradually diminish. In southern Chile a part of the cordillera descends beneath the sea, forming innumerable islands with steep slopes. The Andes have been deeply carved by glaciers, particularly in the south. Glaciers still occupy some 1,900 square miles (4,900 square kilometres), constituting a huge ice cap with long terminal tongues running into lakes or into the sea.

The Andes are studded with numerous volcanoes that form part of the Circum-Pacific volcanic chain, often called the Ring of Fire. Earthquakes are frequent. Almost every major city has been devastated at least once by earthquake, even along the coastal plains, where clear signs of recent vertical movement are visible.

The plateaus. To the north and east, the Guiana and Brazilian highlands consist of ancient crystalline rocks greatly worn through prolonged erosion. The Guiana Highlands are mostly below elevations of 1,000 feet, with small rises separated by marshy depressions. Occasional dome-shaped granitic inselbergs (steep-sided residual hills)—some 2,000 feet in elevation—surmount the landscape. The southern edge rises abruptly to a series of mountain chains and high tablelands (*tepuis*), in which the highest peak is Mount Roraima (9,094 feet).

The Brazilian Highlands

Covering an area of about 580,000 square miles, the Brazilian Highlands (also called the Brazilian Plateau) rise to an average elevation of about 3,000 feet and are crowned by numerous ranges of hills. Included in this region is Bandeira Peak (9,482 feet), one of the highest points in Brazil. The São Francisco River, draining a large basin to the east, has cut deeply into the highlands. In the north the highlands slope gently to the sea, but in the east they drop abruptly, as much as 2,600 feet within a few miles. Skirting their southern edge, the Serra do Mar has summits of more than 7,000 feet in elevation. The sea has partly invaded the original coastal ranges and

Tablelands (*tepuis*) rising behind Hacha Falls on the Carrao River, in the Guiana Highlands of eastern Venezuela.

Physiographic regions of South America.
Adapted from *Odyssey World Atlas*, © General Drafting Co. Inc

formed Guanabara Bay, which includes the harbour for Rio de Janeiro. Nearby are such steep-sided rocky masses as Sugar Loaf (Portuguese: Pão de Açucar; 1,296 feet) and Corcovado Peak (2,309 feet), which rise dramatically from the sea.

In the far south, Patagonia constitutes a series of vast tablelands that rise, terracelike, from the Atlantic to the Andes and are covered with rounded pebbles and crumbling sandstones. Geologically recent volcanic eruptions have spread sheets of basaltic lava over large parts of southern Patagonia and have dotted the sedimentary plateaus with volcanic cones.

The lowlands. The Orinoco River basin is nearly coextensive with the Llanos. It lies between the coastal ranges of the Venezuelan Andes and the Guiana Highlands and is covered with alluvia brought down by the Andean torrents.

The Amazonian depression, the largest river basin in the world, forms an enormous region, bounded by the Andes to the west, the Guiana Highlands to the north, and the Brazilian Highlands to the south. The ancient platform of Precambrian rock underlying the depression is covered with deep layers of alluvial sand and clay, so that it forms an immense plain of low undulations, the general incline being extremely slight. The river port of Iquitos, Peru, which is about 2,500 miles from the Atlantic Ocean, is at an elevation of only 384 feet, while Manaus, Braz., far downstream in the heart of the basin, has an altitude of 144 feet.

The basin of the Paraguay River, between the Bolivian Andes in the west and the Brazilian Highlands in the east, consists of a series of alluvial plains drained by a complex network of rivers interspersed with marshes. To the east, the marshes are called the Pantanal. They are only a few hundred feet above sea level, are subject to annual flooding, and form an immense swamp during the rainy season. The extensive plains west of the river, called the Gran Chaco, generally are arid.

The Pampas of Argentina, covering almost 300,000 square miles, consist of an enormous accumulation of loose sediments brought down from the Andes. These deposits, 1,000 feet deep at Buenos Aires and even deeper in other places, have completely buried the ancient features of the land. The landscape seems perfectly level, although it actually rises imperceptibly toward the west—from near sea level at Buenos Aires to 2,320 feet at Mendoza. Some hills, such as the Córdoba and San Luis ranges, are conspicuous features on the otherwise flat plains.

Detailed discussion of the Andes Mountains, Gran Chaco, and Patagonia can be found in *South American geographic features of special interest* at the end of this article.

DRAINAGE

Rivers. Drainage has been notably affected by the physical dissymmetry of the continent. The major basins lie east of the Andes, and the main rivers flow to the Atlantic Ocean. The four largest drainage systems—the Amazon, Río de la Plata (Paraguay, Paraná, and Uruguay rivers), Orinoco, and São Francisco—cover nearly three-fourths of the continent.

The Amazon basin

By far the largest system is formed by the Amazon River. Stretching some 4,000 miles across equatorial South America, the Amazon is second in length only to the Nile. The volume of water it carries surpasses that of all other rivers, constituting one-fifth of the total flowing fresh water of the world. About 6,350,000 cubic feet (179,800 cubic metres) of water per second is emptied into the Atlantic by the Amazon, which is more than 10 times the outflow of the Mississippi River. The Amazon drains some 2,722,000 square miles—about two-fifths of South America—and has more than 1,000 tributaries, several of which are more than 1,000 miles long. Rising in the central Peruvian Andes, it is named the Marañón in its upper course; after being joined by several rivers—including the Ucayali, from which the Amazon's length is measured—it escapes from the Andes through narrow canyons. Near Manaus, it is joined by the Negro River, which drains much of northern Brazil. The Amazon, then at full strength, winds through the low plains to pass between the Guiana Highlands and Brazilian Highlands before emptying into the Atlantic.

The second most important drainage system, estimated to cover at least 1,600,000 square miles, is formed by the Paraguay, Paraná, and Uruguay rivers. These empty into the Río de la Plata, which actually is an estuary or gulf and not a river. About 2,800,000 cubic feet of water per second discharge from the common mouth of these rivers, an outflow second only to that of the Amazon. The Paraguay River, with a length of 1,584 miles, rises in the Bolivian hills and empties into the Paraná River. The Paraguay is a river of the plains, flowing across a wide stretch of marshes (the Pantanal) in its middle course; its lower course, however, is drier. The Paraná has a total length of 3,032 miles; its upper course (generally called the Alto Paraná) flows mainly across high plateaus before its confluence with the Paraguay, after which the river flows through a broad floodplain before emptying into the Río de la Plata. The Uruguay River, at 990 miles, is the shortest of the three; it flows east of the Paraná before discharging into the Paraná delta.

The Orinoco River basin is the third largest drainage system, covering about 366,000 square miles. With a length of some 1,700 miles, the river first flows west and then north, plunging down a series of steep slopes. It then flows northeast and east along the edge of the Llanos, an almost flat basin that stretches westward to the Andes. Near the ocean, the Orinoco divides into a series of distributaries to form its delta. A unique feature of the Orinoco system is Casiquiare River, which allows water from the Orinoco to enter the Amazon system.

The basin of the São Francisco River, encompassing some 244,000 square miles, is South America's fourth largest drainage system. The river, which flows entirely within Brazil, has a total length of 1,811 miles. It rises in the state of Minas Gerais and flows northward for 1,000 miles before curving eastward to the Atlantic. The São Francisco has been an important artery of communication since colonial times and now is the location of several large hydroelectric projects.

The remainder of the continent's Atlantic-flowing rivers are of much less importance. Among the largest are the Magdalena in Colombia (navigable in its lower section) and the Essequibo, Maroni, and Oyapock in the Guianas.

Drainage to the Pacific is different because of the close proximity of the Andes to the Pacific coast and the scarcity of rainfall from southern Ecuador to central Chile. Consequently, the rivers are short, and few convey any large quantity of water. The major rivers are the Guayas in Ecuador, on which the port of Guayaquil is located, and the Santa in Peru. Some torrents have had great importance since early times, when good water management facilitated the development of ancient civilizations. In central Chile the valleys of the Aconcagua and Bío-Bío rivers have remained fertile agricultural regions.

Lakes. Most of South America's important lakes are in the Andes or their foothills. Because of the chain's complex topography, water has accumulated in closed basins to form natural reservoirs. Among permanent lakes, the largest is Lake Titicaca, which lies at an altitude of 12,500 feet between Peru and Bolivia. The lake is 120 miles long and up to 50 miles wide, although it was much more extensive in the past. Lake Junín in central Peru; Lake Sarococha, also in Peru, between Puno and Arequipa; and Lake Poopó in Bolivia also rank among the larger Andean lakes. They exhibit uniform physical conditions throughout the year, in terms of temperature and percentage of dissolved gases. In addition, they remain ice-free up to an altitude of 16,000 feet, and, as a result, the climate of their shores is temperate.

The Andean lakes

Another type of lake is found in Patagonia where, in the wake of melting glaciers, lakes formed downslope in

An Aymara Indian poling a reed boat on Lake Titicaca, Bolivia. The Cordillera Real in the Bolivian Andes rises in the background.

natural basins. Among these are Lakes Buenos Aires, Argentino, and Nahuel Huapí. Their eastern parts, which stretch to the end of the Argentine plateau, generally have gently sloping banks bordered by low mountains, while their western parts form a series of narrow, fjordlike arms that lie between steep slopes.

Marshes and swamps. Vast marshes are found in depressions in many parts of the continent. One of the widest marshy areas is the Pantanal, in the middle course of the Paraguay River; it is subject to flooding in December, reaching its highest watermark in June, when it becomes an immense swamp. Swamps of another type occur in the rain forests, mostly in the Amazon basin and in northwestern Colombia. In some places the ground is inundated throughout the year, whereas in other areas swampy conditions occur only at the time of the annual flood. Finally, wide, marshy areas border the mouths of the Orinoco and Amazon rivers, and mangrove swamps of various types are found along the lower river valleys and coasts from southern Ecuador northward, less continuously along the Caribbean coast, and south along the Atlantic coast to southeastern Brazil.

Detailed discussion of the Amazon River basin, the Orinoco River basin, the Río de la Plata system, and the São Francisco River can be found in *South American geographic features of special interest* at the end of this article.

SOILS

Because of its geologic history, topography, climate, and vegetation, some 27 distinct soil regions can be distinguished in South America, many of which have several subdivisions. The soils constitute three major groupings that correspond to the continent's three primary geographic components—the lowlands, the highlands, and the Andes.

Low natural fertility is a conspicuous feature of soils in the tropical regions of South America. About half of the continent's soils consist either of unconsolidated and nutrient-poor sediments (*e.g.*, kaolins [china clays] and quartz sands) found in river basins, latosols (red soils leached of silica and containing residual concentrations of iron and aluminum sesquioxides), red-yellow podzols (acidic soils with a bleached upper horizon, or layer, that are low in lime), and regosols (azonal soils consisting mainly of imperfectly consolidated material and having a complex morphology). About one-fifth of the continent is covered by desert soils of various types in which agriculture is risky without irrigation. Other regions, representing about 10 percent of the total area, are poorly drained, the soils being either gleys (clayey soils in which the substrate is bluish gray, generally sticky, and often structureless because of excessive moisture), groundwater laterites, grumosols (soils with a high content of expanding clays), or

Adapted from R. Ganssen and F. Hadrich (eds.), *Atlas zur Bodenkunde*

Soils of South America.

planosols (a type of soil found in humid climates in which soluble salts and minerals are leached out of the upper layers and are cemented or compacted at a lower level). In the Andes, slopes are often steep, and lithosols (shallow soils consisting of imperfectly weathered rock fragments) abound, accounting for another 10 percent of the continent's surface. In the inter-Andean valleys and on some of the foothills, nevertheless, eutrophic soils (deposited by lakes, and containing much nutrient matter, but often shallow and subject to seasonal oxygen deficiency) can be found.

The fertile soil regions

Fertile soils, therefore, extend over only about 10 percent of the surface of South America. The most important of these are brunizems (deep, dark-coloured prairie soils, developed from wind-deposited loess), chestnut soils, and ferruginous tropical soils. On the low coastal ranges, in the foothills of the western Andes, and on the nearby plains and terraces of Colombia and Ecuador, the soils consist mainly of red-yellow latosols, podzols, and alluvial soils. Soils in southern Brazil and Uruguay consist of brunizems, reddish prairie soils, and planosols. The Argentine Pampas, the largest fertile area on the continent, is uniformly covered with the so-called pampean loess, which is calcareous, rich in minerals, and partly of volcanic origin. Less rich soils are found in the uplands of northeastern and central Brazil, consisting mainly of sandy regosols in the north and red latosols in the south.

The agricultural development of South America closely reflects the distribution of soils according to their fertility. It is mostly confined to the eastern mid-latitude plains, in which is concentrated the production of cereal grains and cattle grazing; to the subtropical and temperate parts of the Andes, from Colombia to Chile, where grazing takes place and a variety of crops are cultivated; and to eastern and southeastern Brazil, where coffee, cacao, soybeans, and sugarcane are grown, while the interior plateaus are devoted to cattle grazing.

Soil erosion has ravaged a large part of the continent. According to some estimates, in several countries half or more of the presently arable land has been severely damaged or ruined by poor land management. In the Andes, land that once produced high yields of wheat is now abandoned. Mountain forests are still cleared for cattle grazing and cultivation, which greatly accelerates erosion and ruins the soil of the region for years thereafter. Soil damage has been less severe in areas of relatively flat terrain. Campaigns for soil conservation or restoration have been implemented in most countries.

CLIMATE

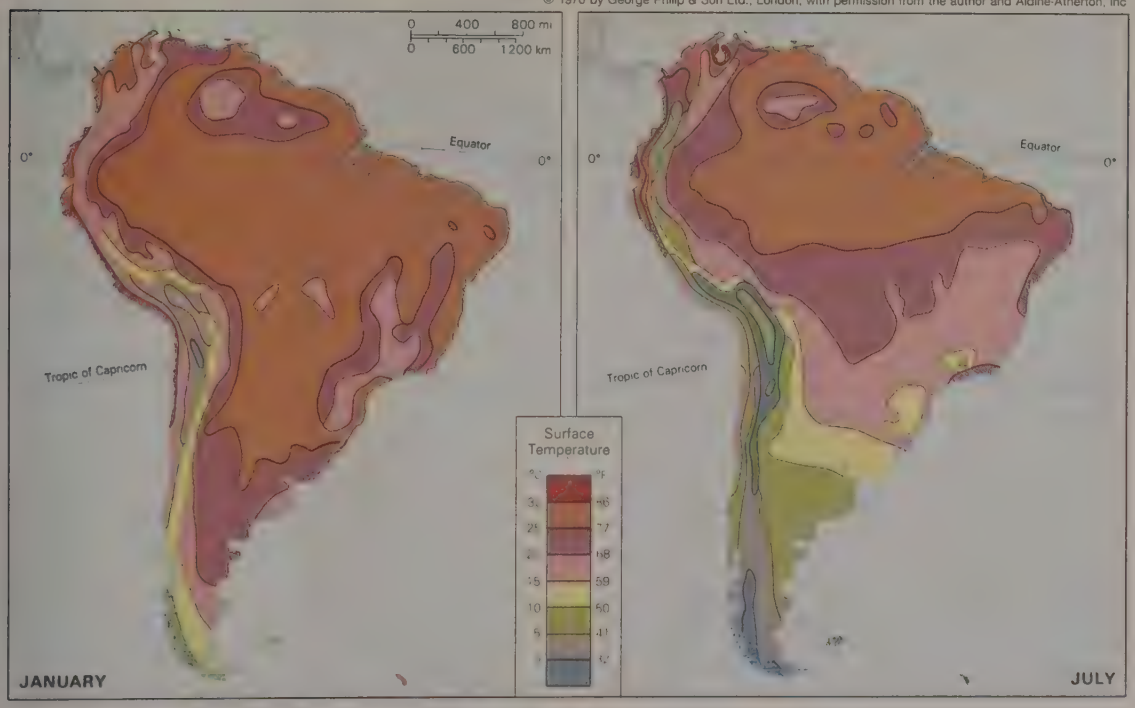
South America extends over a wide range of latitude, and encompasses a great variety of climates. Although it is the only continent (other than Antarctica) to reach such a high southern latitude, South America's broadest extent is in the equatorial zone, so that tropical conditions prevail over half of the continent.

Factors influencing climate. Three principal factors control the broad features of South America's climate. The first and most important of these are the subtropical high-pressure air masses over the South Atlantic and South Pacific and their seasonal shifts in position, which determine both large-scale patterns of wind circulation and the location of the rain-bearing intertropical convergence zone (ITCZ). The second is the presence of cold ocean currents along the continent's western side, which affect both air temperatures and precipitation along the Pacific coast. Finally, the orographic barrier of the Andes produces a large rain shadow over much of the southern part of the continent.

Air circulation patterns

The South Atlantic and South Pacific high-pressure cells take the form of great semipermanent anticyclones (centres of high atmospheric pressure around which winds circulate), the positions and mean intensities of which change with the seasonal north-south migration of the Sun. The eastern part of the South Pacific anticyclone dominates the climate of most of South America's west coast, producing stable, subsiding air that yields minimal precipitation. The cold Peru (Humboldt) Current, flowing northward along the coast from southern Chile to the Equator, cools and further stabilizes the Pacific air that invades the continent. One of the world's driest regions, the Atacama Desert along the northern coast of Chile, results from these conditions. The east coast (north of Patagonia), by contrast, receives greater amounts of precipitation from the humid winds emanating from the South Atlantic high.

The ITCZ is responsible for the seasonal character of precipitation in South America's extensive tropical wet-dry climatic zone. In this region, low-pressure cells persist between the subtropical anticyclones of the Northern and Southern hemispheres, and the trade winds of both hemispheres converge. A migrating zone of unstable atmospheric conditions results, bringing periods of prolonged,



Average temperatures for January (left) and July (right) for South America.

abundant precipitation. The ITCZ is a product of the subtropical highs, and it also follows the annual migration of the Sun. Thus, the ITCZ reaches its most northerly position during the Southern Hemisphere's winter, which is the driest period for most of tropical South America.

The southern portion of the continent is unaffected by the ITCZ and falls instead under the influence of the mid-latitude westerlies, which are particularly strong in the Southern Hemisphere because of the large extent of ocean area there and the unimpeded air flow this allows. As the westerlies rise over the Andes, most of their moisture is lost in orographic precipitation. A typical rain shadow develops on the lee side of the chain, as in the vast desert and semidesert region of Patagonia.

Finally, altitude is an important factor. Vertical climatic zones, ranging from humid tropical at lower elevations to Alpine and Arctic on the high peaks, are particularly well-defined in the Andes.

Climatic regions. South America can be divided into four major climatic regions—tropical, temperate, arid, and cold—their parameters determined by the factors described above.

Tropical climates. Among the tropical climates, the tropical rainy, or rain forest, type occurs on the Pacific coast of Colombia, in the Amazon basin, on the coast of the Guianas, and on part of the coast of Brazil. The average daily temperature is about 86° F (30° C), with monthly and annual variations of less than about 5° F (3° C). Heavy rainfall, well-distributed throughout the year, averages about 98 inches (2,500 millimetres) annually in Belém (Brazil), about 108 inches in Iquitos (Peru), and 71 inches in Manaus (Brazil). The Chocó region of Colombia—one of the wettest areas in the world—receives in excess of 400 inches, and it rains more than 300 days per year. In the Amazon region, rains do not fall evenly over the basin. The southern part receives most of its rainfall during the Southern Hemisphere summer (October to April), while the northern part has its rainy season during the Northern Hemisphere summer (May to September). The "dry" season is neither lengthy nor noticeable; humidity is always high.

Tropical wet-dry climatic regions

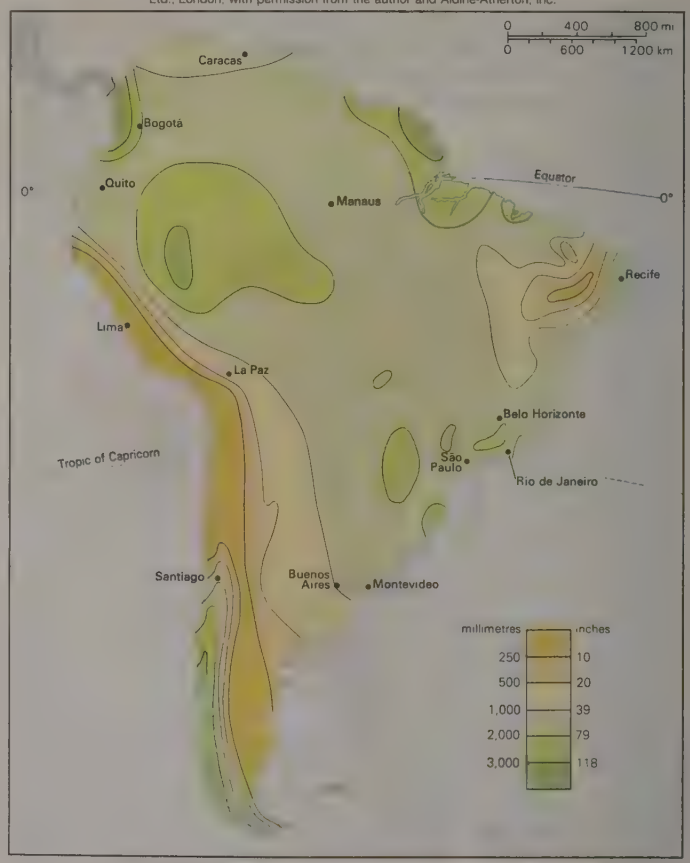
The second type of tropical climate—the tropical wet-dry, or savanna (grassy parkland), type—is characterized by high temperatures (all monthly means above 64° F, or 18° C) but receives less precipitation and experiences a prolonged dry season. It is found around the tropical-rainy belt, in the Orinoco basin, on the Brazilian Highlands, and in part of western Ecuador. Temperatures are still high and annual variations small, but daily temperature extremes are greater, typically ranging from a low of 65° F (18° C) to a high of 95° F (35° C).

Temperate climates. The so-called temperate climates actually have a greater range of temperatures than do the tropical climates and may include extreme climatic variations. These climates, characterized by lower winter temperatures, are found south of the Tropic of Capricorn (in Paraguay, parts of Bolivia, Brazil, Argentina, and Chile) and in the Andes. On the Atlantic side, temperatures in the warmest month average 77° F (25° C), but cold-month averages vary from 63° F (17° C) in the north (Asunción, Paraguay) to 50° F (10° C) in Buenos Aires, Arg. Rainfall exceeds 1.5 inches each month in the east but decreases to the west. In central Chile, between latitudes 32° and 38° S, the climatic features are similar to those of the Mediterranean, with mild winters and winter rains; summers, however, are cooler (69° F, or 21° C, in Santiago, Chile, in January—9° F, or 5° C, cooler than in Mediterranean locations). In southern Chile, winter temperatures are lower but not as low as the latitudes would indicate. The islands and channels have a relatively uniform climate throughout the year, and winters are much less severe than in Labrador, for example, which is at a comparable latitude and maritime location north of the Equator. The presence of glaciers is the result of cold, snowy winters and cool, cloudy summers during which ice does not completely melt. Rainfall is abundant (102 inches in Valdivia, Chile, and probably twice this figure on the western slopes of the mountains), and the coast is one of the wettest regions in South America. A short distance inland, however, after passing into the lee of the Andes, rainfall decreases considerably (20 inches at Ushuaia, Arg.). Thus, in Patagonia an unusual situation exists in which these variations in rainfall result in greater differences in climate from west to east than from north to south.

Arid climates. Arid climates are found in four areas. Patagonia and northwestern Argentina constitute the largest of these. Rainfall is low, only about four inches in San Juan in the north and about seven inches farther south in Neuquén. The annual range in average monthly temperatures in Patagonia—the greatest in South America—is more than 36° F (20° C), the result of warm summers and cold winters. Another arid zone, the Atacama Desert, is found in a narrow strip along the Pacific coast between latitudes 5° and 31° S. The cold seas associated with the Peru Current and the proximity of the high Andes produce an inversion of normal atmospheric temperatures, as air in contact with the water cools more rapidly than the upper strata of air. The result is a nearly continuous layer of stratus clouds about 1,200 feet thick, lying at altitudes varying between about 1,000 and 3,000 feet, that prevents air near the ground from being warmed. Temperatures, consequently, are relatively low: Lima has an average annual temperature of 64° F (18° C), ranging from about 72° F (22° C) in February to about 59° F (15° C) in August. The coast of Peru thus is the cloudiest—and one of the driest—deserts in the world, with no sunshine for at least six months of the year. It almost never rains, except under abnormal circumstances, but condensation of fog (called *garúa* by the Peruvians) provides a limited amount of moisture. Another desert extends from northeastern Colombia to Venezuela, covering a zone where rainfall is scarce and droughts are prolonged.

The coast of Peru

Finally, an arid zone occurs in northeastern Brazil, between the Parnaíba and São Francisco rivers. The interior highlands act as a wedge separating the sea winds from the northeast and those from the southeast, which carry their moisture beyond the region. Average annual rainfall is less



Average annual precipitation in South America.

than four inches, and the dry season may last as long as seven months. The worst feature of the area's climate is the irregularity of the rainfall, as a result of which severe droughts plague the region.

Cold climates. Areas where average annual temperatures are less than 50° F (10° C) are characterized as having cold climates. These occur in the southernmost parts of Argentina and Chile and in the high Andes above about 11,500 feet. Mean temperatures are relatively low, but daily variations are wide. There is a marked difference in humidity between the northern and southern parts of the upper Andean zone. In Colombia and Ecuador the climate at such altitudes is cool and damp. Temperatures, always low, may on the average vary daily from 54° F (12° C) during the daytime to 28° F (−2° C) at night. Rainfall generally is high and well distributed throughout much of the year, although most of the Ecuadorian Andes have a dry period from June to August. Clouds and mist are dense in much of the region, and sunlight penetrates only for short periods. From central Peru to Bolivia and Chile, temperatures are still lower. Near Lake Titicaca, the average annual temperature is only about 34° to 36° F (1° to 2° C); November is the warmest month, with an average temperature of 41° F (5° C), while the coldest month, July, has an average of 28° F (−2° C). Daily variations are considerable; typically, a daytime maximum of 68° F (20° C) may drop to a nocturnal minimum of 5° F (−15° C). Annual rainfall varies from 24 to 56 inches but is concentrated during the southern summer. The dry season is long and characteristically accompanied by drought. Winds are continuous and often violent, accelerating the coldness and the dryness of the climate, which produces a harsh and hostile environment. As in any mountainous region, climate varies largely according to local conditions.

(J.P.D./C.W.M.)

PLANT LIFE

South America, with its distinctive plant and animal life, covers most of the biotic region called the Neotropics (the faunal realm also is called the Neogaean). This region extends southward from the Tropic of Cancer to include all of Central and South America—even that southern portion that does not have a tropical climate. With Australia it is, because of its isolation from the rest of the world during the Tertiary Period (66.4 to 1.6 million years ago), the landmass with the strongest biological originality.

Affinities with other continents

Vegetation and animal life manifest some relationships with other continents as a result of past geologic events. Ancient groups of plants and animals including mollusks, chilopods, some fishes, reptiles, and amphibians show

Vegetation zones of South America.

affinities with the animal life of Africa, Australia, and New Zealand. More recent elements, mostly vertebrates, migrated from or through North America.

The pattern of distribution within the continent is complex because of the variety of climatic and ecological zones. The northern tropical regions are the richest in diversity, while the southern part and the western Andean highlands are much impoverished, despite some differentiation.

Vegetation zones. The proportion of endemic plants in South America is very high, even at the family level. Among angiosperms (plants having seeds enclosed in an ovary) no fewer than 25 families and 3,500 genera are endemic to the tropical and temperate zones. Others are related to African plants or belong to southern plant groups also distributed in southern Africa and in Australasia. Vegetation is by no means uniform throughout the continent; its distribution is determined by climatic, geographic, soil, and sometimes anthropic (human-related) differences.

Tropical and subtropical rain forests. Rain forest covers the largest part of the Amazon region, most of the Guianas, southern and eastern Venezuela, the Atlantic slopes of the Brazilian Highlands, and the Pacific coast of Colombia and northern Ecuador. The huge Amazon region is the largest and probably the oldest forest area in the world; it also ascends the slopes of the Andes until it merges with subtropical and temperate rain forest. On its southern border it merges with the woodlands of the Brazilian state of Mato Grosso, with galleries of its trees extending along the rivers. Consisting of enormous trees, some exceeding a height of 300 feet, the rain forest is composed of an almost incredible number of species growing side by side in the greatest profusion and arranged in different strata. In the region of Manaus, Braz., for example, 1,652 plants belonging to 107 species in 37 different families were found on about 1,900 square feet. Amazonian trees number about 2,500 species.

The forests of the *igapós* (swamps, where the ground is inundated or very marshy throughout the year) cover the lowlands. Characteristic trees are, among others, *jacareúbas* (*Calophyllum brasiliense*), which is a tall tree with hard reddish brown wood used for heavy construction, *araparis* (*Macrolobium acaciaefolium*), *abiuranas* (*Lucuma* species), *piranheiras* (*Piranhea trifoliata*), and *louros-do-igapo* (*Nectandra amazonum*). Undergrowth is dense.

The *várzea* regions are those that are inundated only at the annual flood. Trees are higher and quite diversified; they include *oeiranas* (*Alchornea castaneifolia*), a euphorbia (*i.e.*, characterized by a milky juice), and the trumpet tree (*Cecropia peltata*), a rapid-growing tree of the mulberry family with a light wood. Palms and hevea (a kind of wild rubber plant) grow in these forests. The forests of the nonswampy areas are rich in hardwoods, of which acapu (a tree with dark brown wood), *pau-amarelo* (*Euxylophora paraensis*), *pau-santo* (*Zollernia paraensis*), massaranduba (a Brazilian tree with light reddish brown wood), jarana (a tree with hard, heavy, durable wood), and matamata (a tree with hard, heavy, durable wood, used for pilings) are the best known. Hevea and the Brazil nut tree (*Bertholletia excelsa*) are characteristic of these forests where spiny palms cover the ground.

Epiphytes (nonparasitic plants that grow on other plants, deriving moisture and nutrients from rain and air) are numerous, mostly Bromeliaceae (a family having spiny leaves), orchids, and ferns. Lianas abound, particularly in drier forests. Epiphytes

Tropical deciduous forests. These forests, dominated by trees of moderate height, notably of leguminous species, are found widely dispersed through northern South America, where the climate is characterized by a prolonged dry season, notably in Venezuela, Colombia, and the Brazilian Highlands.

Caatinga. Caatinga (white forest) refers to the generally stunted, somewhat sparse, and often thorny vegetation of the dry interior of northeastern Brazil. Trees, leafless for long periods and able to resist drought, also are characteristic, particularly in the basin of the São Francisco River. Dominant species are leguminous trees, particularly *catingueiras* (*Caesalpinia*), *juremas* (*Mimosa*), and *joazeiros* (*Zizyphus joaseiro*), members of the Euphorbiaceae (spurge) family, and Bombacaceae (a family of tropical trees with palmate leaves and large, dry or fleshy fruit). Undergrowth consists of thickets, bromeliads (plants with basal, often spiny leaves), and innumerable cacti, among which is the *xiquexique* (*Cereus gounellei*), the complicated intertwinings of which cover the soil. Where more water is available, caatinga species may grow 30 feet high and form impenetrable thickets.

South Brazilian forests. These parklike forests, sometimes very dense but interspersed with savanna, occupy vast expanses from the border of the Amazonian rain forest to the marshes of the upper Paraguay River. The typical landscape is a grassland strewn with smaller trees. In effect, it includes a mosaic of associations, from hygrophilous (living or growing in moist places) to xerophilous (adapted to dry conditions) forests and even desert.

Notable is the forest region of araucaria, or Paraná pine (*Araucaria angustifolia*), which covers a vast area between the Paraná River and the Atlantic Ocean, stretching from Curitiba, Braz., to northern Argentina. Araucarias (which are not true pines) dominate a dense forest of numerous species including hardwoods, yellowwood (Podocarpus), and the South American holly (*Ilex paraguariensis*), from which the beverage maté is made.

Xerophytic associations. Thickets of small trees and

Rain-forest vegetation along the northern coast of Ecuador.

Caatinga vegetation in the dry interior of northeastern Brazil.

shrubs, often thorny, among which species of *Prosopis*, *Acacia*, and *Mimosa* predominate, cover regions that experience alternate dry and relatively wet seasons; these regions particularly include coastal Venezuela, northeastern Colombia, southwestern Ecuador, and northern Peru.

In Peru this association merges with the desert, which extends along the coast to northern Chile, having a width of from 50 to 100 miles. Only a few shrubs and some terrestrial (as distinct from epiphytic) bromeliads grow in this area, which becomes greener only in the Andean foothills, where cacti and other xerophytic plants are to be found, and along the valleys of rivers flowing down from the Andes. In some areas, called lomas in Peru, winter mists bring some humidity, and a specialized type of vegetation, consisting of annual and bulbous plants, grows for a short period.

Lomas vegetation

Subantarctic beech forests. Temperate rain forests—similar to those found in British Columbia and in the northwestern United States—grow in southern Chile at low and moderate altitudes, thanks to abundant rainfall. The most typical trees belong to the genus *Nothofagus* (timber trees found in the cooler parts of the Southern Hemisphere), the northern species of which are evergreen and the southern species deciduous. Various conifers, notably the alerce and araucarias, mingle with the leafy trees. A dense undergrowth of shrubs, lianas, bamboos, ferns, mosses, and epiphytes grow in the northern districts but disappear toward latitude 49° S. In southern Chilean Patagonia, forests consist of twisted, creeping trees merging into a kind of heath.

Savannas. True savannas are found mostly in the Llanos of Venezuela and northeastern Colombia. This vast area is covered with grasses and sedges, but trees (mainly palms) also are found, especially along streams.

Pampas. The flat plains called Pampas, which constitute the greater part of eastern Argentina, are covered with grasses. The Pampas originally were covered with trees, but humans have removed the trees and have acclimatized various grass species. Exotic pines, eucalyptus, oaks, and poplars constitute introduced trees. To the south the Pampas merge with the Patagonian steppe, where tussock grasses are mixed with scattered low bushes and spiny plants. The vegetation becomes much poorer to the south.

Mountain vegetation. In the high Andes, a temperate zone extends from the upper limit of the subtropical rain forest (about 6,000 feet) up to the timberline, which reaches an altitude of about 11,000 feet in Ecuador and Peru. This zone is very humid on the Amazonian side, where both the epiphytes and the undergrowth are dense.

The upper zones have a peculiar vegetation that reaches the snow line. In the wet northern Andes in Colombia and Ecuador, Alpine meadows called *páramo* consist of grasses and other herbaceous plants, often with bright flowers; these are surmounted by taller plants, especially the showy frailejones (*Espeletia*), which grow 15 to 20 feet high and are crowned by large bouquets of long hairy leaves.

In the south *páramo* vegetation merges with that of the high plateaus, notably of the Altiplano of southeastern Peru and western Bolivia. Typical vegetation consists of rough grasses, between which grow a variety of herbaceous, cushion, and rosette plants, shrubs, and cacti. Vegetation extends to the limit of permanent snow but becomes scarce at higher altitudes where soil is often barren.

Human influences on vegetation. Human activity has transformed the original vegetation cover to a large extent throughout South America, particularly in forested areas. The forests of eastern Brazil were ravaged in the process of clearing the ground for crops, especially sugarcane. The forests of araucaria in the southern states of Brazil have been rapidly vanishing, as they have been exploited for timber. The slopes of the Andes are so severely deforested that it is not apparent that they once were covered with trees. In the Amazon region, hundreds of square miles of tropical rain forest are cut annually, and the forest no longer is the enormous inviolate mass it once was. In Patagonia the practice of burning to convert remaining patches of forest into pasture steadily increases. Animal herders have severely damaged grasslands through overgrazing, from the Venezuelan Llanos through the high Andes, to Tierra del Fuego. The destruction of habitats continues to accelerate throughout the continent, despite the growing concern of those favouring conservation.

Destruction of the forest cover

ANIMAL LIFE

South American animal life is particularly rich and well diversified as a result of the wide range of habitats. A number of animal groups well distributed over the rest of the world are nevertheless missing because of South America's lengthy isolation. Many animals belong to exclusive groups, and even at the family level the percentage of endemic forms is high. Speciation has reached a higher degree in South America than in other parts of the world.

Principal faunal types. *Fish and bird life.* Freshwater fishes are numerous, with about 2,700 species, though they derive from only a few ancestral groups. Amazonian fishes may approach 1,500 species in number. Among the dominant groups are characins (800 species), which include the flesh-eating piranha; gymnotids, South American cyprinoid fishes that include the electric eel; catfishes; cyprinodonts, a large family of small scaly-headed soft-finned fishes; and cichlids, a family consisting chiefly of fishes that somewhat resemble sunfish.

Birds are represented by 89 families and some 3,000 species—a much higher figure than in Africa or Asia, which justifies the application of the name bird continent to South America. Twenty-five families are endemic to the Neotropical region. Unique birds include rheas (large, tall, flightless birds that resemble ostriches), curassows (large arboreal birds distantly related to the domestic fowl), hoatzins (a brownish crested bird, having claws on the digits of the wing when young), oilbirds, motmots (bright-coloured birds related to kingfishers), jacamars (small, bright-coloured birds), toucans, manakins, and cotingas (related to manakins), and many passerine (perching) birds. Hummingbirds have evolved to fill a variety of habitat niches, with more than 120 species in Ecuador alone. Parrots, pigeons, cuckoos, tyrants (a kind of flycatcher), woodhewers, and orioles are among the dominant groups. Remarkably, the proportion of nonpasserine to passerine birds is greater in South America than in any other part of the world.

Hummingbirds

Mammals. Mammals include types that immigrated to the continent before its complete isolation in the early Tertiary Period. Among these are marsupials (pouched animals), sloths, anteaters, and armadillos; other kinds, now extinct, persisted until the Pleistocene Epoch. More recent elements include hystricomorph rodents (a suborder including porcupines and guinea pigs) and monkeys among the oldest arrivals, and tapirs, deer, carnivores, and bats among the newer arrivals.

Amphibians and reptiles. Amphibians are well represented by caecilians (small wormlike, burrowing amphibians), salamanders, toads, and a number of varieties of frogs, including clawed frogs, the most aquatic of all. The tree frogs, arboreal amphibians, are particularly abundant through the Amazon basin and are very different from their African and Asian counterparts, although the frog faunas of Australia and South America often are strikingly alike. Reptiles include a great variety of turtles and tortoises, crocodiles, caimans (endemic crocodilians), geckos, many iguanas, teiids (a family of mostly tropical American lizards), *Amphisbaena* (a genus of harmless, limbless lizards), and many snakes, including boas, colubrids (a very large family of nonvenomous snakes), coral snakes, and vipers.

Arthropods. Most South American insects, spiders, crabs, centipedes, and millipedes are found nowhere else in the world. Thousands of species, especially insects of the tropical rain forest, have yet to be discovered. South America has the richest array of butterflies of any continent, including the spectacularly coloured members of the Morphidae subfamily; the social insects—termites, ants, wasps, and bees—also are well represented. Many of the best-known arthropods (*e.g.,* mosquitoes, sand flies, and kissing bugs) are responsible for the transmission of human diseases.

Ecological communities. Animals are distributed according to the pattern of vegetation zones, and several well-defined communities can be distinguished. They include regions as diverse as the Amazonian forests and the high Andes.

The Amazonian and Guianan forests. The most diverse community is found in the Amazonian and Guianan forests, where the abundance of water and trees makes life easy. Rivers are the realm of large numbers of invertebrates and fishes, such as *pacu* (*Metynnis*), which is a big brownish flat fish, the meat of which is highly valued; *coumarou* (*Curimato*), which is a toothless vegetarian fish resembling the marine mullet; electric eel (*Electrophorus electricus*); pirarucu (*Arapaima gigas*), which can attain a length of 15 feet and a weight of 200 pounds; and piranha, having teeth so sharp that they can cut through flesh like a razor; as well as a wealth of small fishes, many of which are vividly coloured. The manatees (a chiefly tropical, aquatic, herbivorous animal with a broad tail) and the *inia,* a primitive dolphin, frequent the larger rivers of the region. The Amazon river turtle (*Podocnemis expansa*) persists despite intense exploitation. Crocodiles and caimans inhabit the main waterways.

Arboreal adaptions

Amazonian forests constitute an environment to which most animals responded by becoming arboreal. Tree frogs can move across the surface of the leaves thanks to adhesive pads on their feet; lizards have very elongated fingers; monkeys, *sarigues* (a close relative of the opossum), and kinkajous (a nocturnal, carnivorous mammal) have prehensile tails. Birds are numerous and, because of the enormous amount and variety of available foods, well diversified. Antbirds, tyrants, cotingas, *tangaras* (brilliantly coloured birds related to the finches), hummingbirds, toucans, woodpeckers, barbets (loud-voiced tropical birds closely related to honey guides), parrots, and tinamous (quaillike terrestrial birds) are the dominant groups. Many of them never leave the forest canopy, where they display their brilliant colours, which contrast with the more modest plumage of those birds that live in the undergrowth. Mammals are represented by a number of terrestrial or half-aquatic species, such as very small deer; some large rodents (including the agouti, paca, and capybara); tapirs; and carnivores (including the jaguar). Primates range from pygmy marmosets to larger durukulis (small, round-headed, stocky-bodied, bushy-tailed monkeys), woolly monkeys, spider monkeys, and howler monkeys. Bats also are numerous, including species that feed on fruit and others that eat fish. Sloths feed on the leaves of certain trees, among whose branches they remain much of their lives. The predators are represented by carnivorous mammals, a series of snakes—including anacondas and boas—and large raptors (birds of prey), such as the harpy eagle (*Harpia harpyja*), the most powerful bird of prey to be found in the world. Insects, including a great variety of butterflies and ants, are innumerable.

East-central plateaus and lowlands. The Brazilian Highlands have an impoverished animal life, from which species that are strictly adapted to the dense forest are excluded. The plains of Uruguay and the Gran Chaco have a varied animal life that includes some particular species, such as the maned wolf. The marshes are inhabited by a wealth of waterfowl, as well as by a species of lungfish (*Lepidosiren paradoxa*) that is related to its African and Australian counterparts.

The Argentinian Pampas. The Pampas of Argentina are inhabited by only a limited number of indigenous animals. Among the birds are rheas and a series of smaller birds, including the popular ovenbird (*Furnarius rufus*), the name of which comes from its globe-shaped nest made of mud. Endemic mammals include the mara (*Dolichotis patagona*), a long-legged, long-eared rodent; the plains viscacha (*Lagostomus*), a burrowing rodent related to the chinchilla; the guanaco (*Lama guanacoe*), a South American mammal related to the camel but resembling a deer; and Pampas deer (*Blastoceros campestris*). The restricted number of the larger herbivorous mammals is quite remarkable and illustrates the scarcity of recent mammalian types in the Neotropical region.

The southern Chilean forests. The forests of southern Chile are inhabited by a specialized animal life, with a high percentage of endemic species. Parakeets and hummingbirds are found as far south as Tierra del Fuego. A marsupial, the rincolesta of Chiloé (*Rhyncholestes raphanurus*), is one of the most primitive mammals still in existence.

The high Andes. The high Andes have an impoverished animal life. Species there have had to adapt to the harsh and cold environment, scanty vegetation, and low oxygen pressure. The great number of lakes in the region has attracted many aquatic birds, including flamingos, which nest up to elevations of 16,000 feet in northern Chile, and amphibians such as the giant toads of Lake Titicaca, which spend their entire life in water. Mammals are represented by the guanaco and vicuña (both wild ruminants related to the domesticated llama), deer, and numerous rodents, including viscachas, chinchillas, and guinea pigs. Predatory species include foxes, pumas, and many birds of prey, notable among which is the condor, the giant of living birds, with a wingspan of more than 10 feet.

Andean mammals

The arid west coast. The limited variety of animals that inhabit the arid coast of Peru and northern Chile is especially striking when compared with the richness of offshore marine life. The cold upwelling water of the Peru Current, rich in salts, are swarming with life, from plankton to fishes, including the Peruvian anchovy (*Engraulis ringens*); these small forms of life provide food for higher levels of the marine community, represented, for example, by sea lions and birds, many of which are endemic to the area. Birdlife includes a penguin, many species of gulls and terns, shearwaters, petrels, cormorants, pelicans, and boobies (a kind of gannet). Three kinds of these birds—the guanay (or Peruvian cormorant; *Phalacrocorax bougainvillii*), the variegated booby (*Sula variegata*), and the brown pelican (*Pelecanus occidentalis*)—nest by the millions on small islands off the coast, where their droppings accumulate to form guano, a highly prized source of fertilizer.

Human influences on wildlife. Overhunting and habitat destruction have seriously depleted populations of wild animals in much of South America. Almost all wild species are less abundant than they were before the mid-20th century, and some are threatened with extinction. Laws designed to protect wildlife frequently are not observed. In addition, many rural people, especially in tropical-forest areas, still depend on game as a source of food; and the sale of live animals for pets or laboratory use has further depleted stocks. Populations of animals not considered economically valuable have been reduced as their forest habitats have been removed.

Nature reserves, established to protect animals in their habitat, are now found in most South American countries. Argentina pioneered wildlife protection on the continent by creating Nahuel Huapí National Park. At Iguaçu

(Iguazú) Falls—located on the Iguaçu River on the border between Brazil and Argentina, just before the confluence of the Iguaçu and Paraná rivers—two national parks, one in each country, protect wildlife and the surrounding rain forest. Manu National Park in southeastern Peru protects one of the richest collections of plant and animal life in the Amazon basin, including more than 1,000 species of birds. Venezuela's effort to protect habitats led to the establishment (1962) of Canaima National Park in the Guiana Highlands, which with an area of nearly 11,600 square miles is the largest park on the continent. Overall, South America has about 58,000 square miles of parks, but the inviolability of many of these sanctuaries against the pressures of economic development has not been clearly established in all countries. (J.P.D./D.W.Ga.)

The people

ETHNIC ORIGINS AND MIGRATIONS

Four components of the population

Four main components have contributed to the present-day population of South America—American Indians (Amerindians), who were the pre-Columbian inhabitants; Iberians (Spanish and Portuguese who conquered and dominated the continent until the beginning of the 19th century); Africans, imported as slaves by the colonizers; and, finally, postindependence immigrants from overseas, mostly Italy and Germany but also Lebanon, South Asia, and Japan.

Indians. Before the beginning of the epoch of European exploration and conquest in the early 16th century, South America was almost completely occupied by diverse peoples. Nearly all of these cultural groups practiced agriculture, and most exhibited an extraordinary understanding of their physical environment that had been developed over thousands of years. Although areas such as deserts, mountain peaks, and tropical rain forests appeared to be uninhabited, most of these places were occupied at least occasionally. The societies with the greatest complexity of social organization and densest population tended to be located along the Pacific coast, in the adjacent Andes, and along the major rivers of the Amazon basin. Less complex societies were located away from the rivers and mountains, and nomadic hunting groups were found in the Pampas, Patagonia, and southern Chile.

Agriculture-based village culture and social organization came first to the tropical lowlands of the Amazon basin and valleys of coastal Ecuador and Colombia (*c.* 3000 BC). This culture included religious temple-mound complexes, fine ceramics (based partly on earlier technology for making fire-engraved containers out of bottle gourds), and farming such crops as cassava (manioc) and corn (maize) on periodically flooded plains and levees. These areas eventually became organized into complex chiefdoms containing dense populations, supported in some cases by raised fields—broad planting surfaces separated by ditches that enhanced the fertility of the soil while limiting the possibility of fungal diseases and waterlogging.

The practice of agriculture spread to the desert coast of Peru and Chile and then into the higher elevations of the Andes, and new technologies appeared. In coastal areas, elaborate irrigation networks supported ceremonial centres and (later) true cities such as Chan Chan (near present-day Trujillo on the northern coast of Peru), the capital of the Chimú state. Coastal Peruvian and Ecuadorian cultures (such as Moche and Nazca) produced superb ceramic art and finely woven textiles. In coastal Chile the Mapuche (Araucanian) culture effectively occupied its region through farming and hunting.

In the highlands, fertile soils of volcanic ash were cultivated with the digging stick and a type of foot plow called the *chaquitaclla*. Highland soils also were improved by constructing long earthen irrigation canals or (in the Central Andes) some of the world's most elaborate and beautiful stone-walled terracing. In most parts of the Andes, areas of high population density were organized into chiefdoms—such as the Chibcha of Colombia and the mound (tola) builders of Ecuador—led by powerful, paramount lords. Early cities and empires first developed around Huari (Wari) in south-central Peru and Tiahuanaco in western Bolivia, but the last and best-known empire was that of the Inca (Inka). Called Tawantinsuyu, the Inca state expanded from its homeland in the Cuzco Valley of south-central Peru north to what is now southern Colombia and south to the Maule River in central Chile (the northern limit of the Mapuche culture). The Inca easily conquered the desert coastal cultures by threatening their water supplies but never succeeded with the chiefdoms of the Amazon basin and of coastal Ecuador. Thus, when Inca expansion was halted by the Spanish in the 1530s, the empire was long but narrow, confined to the Andes and Pacific coast. This empire did not include all of the advanced agricultural cultures of the continent, which continued north into Venezuela, east into the Amazon basin, and south into the Mapuche area.

The Inca

Certain areas of South America—notably in the more remote parts of the interior, away from the main rivers—were occupied by simpler village cultures based on shifting cultivation, an agricultural practice still used in many of these areas. Nomadic hunter tribes were located in areas of present-day Uruguay and Argentina and in the extreme south (Tierra del Fuego and Cape Horn). Although these cultures appeared to be simple in organization and technology, they were well adapted to hunting wild animals (*e.g.*, the guanaco), fishing, and gathering edible plants in a harsh environment.

The number of Indians at the time of the conquest is uncertain: estimates vary from 8 to 100 million for North, South, and Central America combined (for the Inca, from 3 to 32 million). More recent estimates that put South America's preconquest population at about 20 million seem more realistic.

Equally controversial is the origin of South America's Indians. Most anthropologists believe that they are descended from people who migrated to North America from Asia between about 60,000 and 20,000 years ago, having crossed the Bering Strait separating northeastern Asia and northwestern North America. It is not known when humans first arrived in South America, although it is fairly certain that people were present in Chile by 11,000 BC.

Iberians. Until the end of the era of Iberian domination, only the Spanish and Portuguese were admitted to their South American colonies. The rigid exclusion of all other foreigners had but few exceptions, though a small number of non-Iberian Europeans settled as a result of illegal or tolerated immigration. Most of the Spaniards came from Castile and the southern regions. Little is known about the principal regions from which the Portuguese came. It is estimated that the total number of *licencias* (authorizations to emigrate) granted by Spain was about 150,000 for the entire colonial period, which lasted from the 16th to the 19th century; it is possible that the number of illegal immigrants also approached this number. Of these, no more than two-fifths of the emigrants went to South America. Up to one million Portuguese may have migrated to Brazil, drawn primarily by a gold rush in Minas Gerais in the 18th century.

Africans. A few African servants accompanying the early Spanish or Portuguese explorers were the first slaves to enter the continent. Larger-scale importation of slaves from Africa developed after the slave trade was established early in the 16th century, though reliable quantitative information is lacking. Estimates of the number of Africans brought to South America are four million for Brazil and three million for all of Spanish America, of which most went to areas of present-day Venezuela, Colombia, coastal Ecuador and Peru, and northwestern Argentina; a number also went to the large Spanish colonial cities as urban servants. In addition, many Africans were brought to the British and Dutch Guianas (present-day Guyana and Suriname, respectively). African slaves were considered to be more resistant than American Indians to tropical diseases, especially in plantation areas. Most of the slaves imported into South America came from Portuguese or Spanish trading posts along the west coast of Africa, including areas near present-day Angola. The slave trade ceased in the early 19th century.

The slave trade

Postindependence overseas immigrants. Most of the South American countries gained independence in the

early 19th century, thus bringing an end to the legal exclusion of foreigners. Mass immigration to the continent, however, did not begin until after 1850, acquiring momentum in the last three decades of the century and continuing until 1930, when it decreased abruptly. Some 11 to 12 million people arrived in South America; the great majority of these went to Argentina (more than half) and Brazil (more than one-third). Although many later left, the demographic and sociocultural impact of this influx was tremendous in Argentina and (to a lesser extent) in southern Brazil. Immigration to other countries was numerically insignificant (although socioculturally meaningful), except in Uruguay, where because the preexisting population was not numerous, the proportion of foreign-born was high—about one-fifth in 1908 and even higher in the 19th century. In Argentina the proportion of foreign-born reached nearly one-third of the total population and stayed at that level for many years. In both cases the contribution of post-independence immigration was proportionally much higher than in the United States at the peak of mass immigration.

The great majority of the immigrants were Europeans—Italians (forming nearly half of the immigrants in Argentina, one-third of those in Brazil, and probably the majority of immigrants in Uruguay), the Spaniards (one-third in Argentina), and the Portuguese (nearly one-third in Brazil). Other small but socially relevant immigrant streams arrived from central and eastern Europe. This source of immigration became more important in the 20th century and especially during the 1930s and '40s, when it included more middle-class and educated people, among whom were many Jews and other refugees. After World War II another smaller wave of immigration arrived from Europe (principally Italy and Spain), directed mostly to Venezuela and Argentina.

Asian immigrants

Other important waves of immigration arrived from East and South Asia and the Middle East. Chinese labourers came in the 19th century to help build South American railways and established Chinese districts in such cities as Lima. Labourers from South Asia were brought by the British to Guyana, and similar migrants went to Suriname, supplemented by workers from the East Indies (Indonesia). Lebanese migrated to South America from the Ottoman Empire prior to World War I; known locally (and incorrectly) as "Turks" (*turcos*), these Lebanese became important in commerce and even politics in such cities as Guayaquil, Ecuador. Since World War II, Koreans have migrated to Argentina (under a negotiated treaty) and under less formal conditions to countries as diverse as Paraguay and Ecuador, where they often have become involved in commerce and industry. The largest Asian group by far, however, has been the Japanese. Before World War II large numbers of Japanese settled in Brazil, Bolivia, and Argentina. People of Japanese ancestry now are found primarily in the Brazilian states of São Paulo, Santa Catarina, and Rio Grande do Sul, as well as in Argentina and Peru; and collectively they constitute the largest concentration of ethnic Japanese living outside of Japan.

POPULATION AND ECOLOGICAL DISTRIBUTION

The present population. The present population of South America is the result of four centuries of ethnic mixture among these four components—indigenous peoples, Iberians, Africans, and overseas immigrants—and their various descendants. The mixing process began when the first Iberians reached South America. The previous traditions and basic values and attitudes of the Iberians—coupled with other characteristics of their conquest and colonization—facilitated intermixing not only with the Indians but in general among all the various ethnic groups, although the intensity, extent, and frequency of this mixing varied both among different groups and at different times.

Legal marriage between Iberians and Indians was tolerated, often permitted, and even, in some special circumstances, promoted. It was possible—and in certain epochs easy—to recognize mestizo (generally, mixed European and Indian) children, though frequently a mestizo was considered automatically illegitimate. Social custom did not permit intermarriage between Europeans and Africans and between Africans and Indians and their offspring, although social custom failed to prevent generalized interbreeding. This prolonged process created a great variety of physical types, resulting in the emergence of a complex terminology to describe them.

The more important designations are *mestizo* (called *caboclo* in Brazil), *mulatto* (European-African), and *zambo* (African-Indian). During the postindependence period of European immigration, other national groups contributed to further ethnic mixture. As a result, Argentina and Uruguay completely modified their ethnic composition.

Culture areas. Many South Americans resist recognizing ancestry as being socially significant, especially as language, religion, and other cultural aspects tend to cross ethnic lines. In practice, however, an individual's ethnic background can be a factor in determining social status or economic opportunities. Ethnic distinctions tend to be geographic in nature and can be divided into three broad types of regions—Indian-American (indigenous American), African-American, and European-American—based on a predominant ethnic element in what otherwise are mixed populations. These regions have been defined to a large extent since the end of the colonial period. The people within them, however, are not of uniform ancestry but rather are clustered into different cultural groups.

Three cultural regions

Indian-American regions. Included in this designation are the Amazon and Orinoco basins, Paraguay, northern and central Chile, and the highlands of Venezuela, Colombia, Ecuador, Peru, and Bolivia. These were the areas of pre-Columbian chiefdoms and states at the time of the Spanish conquest. It was possible to found the social and economic organization of the Spanish empire on the relatively advanced institutions of these cultures; at the same time, Indian labour could be easily exploited. The original Indian population suffered from what has been regarded as a demographic disaster—during the first century of Spanish domination the Indian population declined up to 95 percent in some areas—but a substantial number of Indians survived, and mixing between Indians and Europeans often was intense. Indian populations in the highlands began to recover during the 18th century. The European component of the population was confined to towns and cities in colonial times, followed by more recent migrations to areas of the Amazon basin during the rubber boom of the early 20th century. African populations were limited to plantation zones in warmer mountain valleys. In Chile a pronounced shift toward identification with European culture took place, even though the population had a substantial proportion of Indians.

African-American regions. This category includes the coastal areas of the Guianas, Venezuela, Colombia, Ecuador, and Peru; most of eastern Brazil; and northwestern Argentina. Most of these areas are not considered "black" today but rather are identified in regional terms. Thus, coastal Ecuadorians in general are called *montuvios,* coastal Peruvians *criollos,* and northwestern Argentines *criollos* or *gauchos.* Many African-American Colombians who were designated as *mulattos* on colonial censuses were called *mestizos* on later censuses. Changes in African cultural identity were facilitated by the rapid adoption of European religion and language in slave populations.

European-American regions. Most of the people included in this grouping live in a belt extending from southern Chile through Patagonia and the Pampas to southern and southeastern Brazil. Part of this area contained hunter-gatherers at the time of the Spanish conquest, and these peoples strongly resisted European domination until they were decimated by modern warfare in the 19th century. The Brazilian area contained Indians who had been displaced westward by slave raids during colonial times. People of European and Japanese ancestry were drawn into these regions in the 19th and 20th centuries by the possibility of developing forms of commercial agriculture—especially wheat and dairy farms and cattle and sheep ranches—similar to those in their homelands.

Diversity

Linguistic patterns. The linguistic diversity and multiplicity of South America probably is unmatched anywhere else in the world. Thousands of languages and dialects

have been cataloged, including all those existing since the European conquest. Classification systems vary a great deal—from more than 100 "linguistic families" and many unrelated languages at one extreme to extremely simplified schemes at the other. There also is considerable disagreement on the composition of these "stocks" and how many languages should be classified. Most are now extinct, either because the peoples who spoke them have disappeared or because of acculturation into a European language or, in some instances, into another indigenous tongue.

The survival of Indian languages in the Indian-American areas has depended on a variety of factors. Colonial authorities helped spread Quechuan languages (those spoken by the Inca) because they were convenient for missionary activities and for government, and these languages often displaced local indigenous languages. Elsewhere, local languages gave way to new languages such as the *língua-geral* of Brazil (combining Tupí-Guaraní and Portuguese). In many cases populations became bilingual, with an Indian language spoken at home and Spanish used for public transactions; examples include the Spanish-Guaraní speakers of Paraguay and the Quechuan-Spanish speakers throughout the Andes.

The largest surviving indigenous language groups are Quechuan, Aymaran, Tupí-Guaraní, and Mapuche. Quechuan languages are in use primarily in the Andean highlands (southern Colombia to Bolivia) but also in large areas of the Amazon basin and in northwestern Argentina. Quechuan, collectively the third largest language group in South America after Spanish and Portuguese, is not spoken by all Andean highlanders but is limited to certain sharply defined geographic domains. Aymaran languages are spoken in northwestern Bolivia, southeastern Peru, and small areas of northwestern Argentina and northern Chile. Most people in Paraguay speak Spanish and a dialect of Tupí-Guaraní and consider themselves to be mestizo Paraguayans rather than Indians. Mapuche speakers, who constitute the largest Indian population in Chile, are restricted to the south-central part of the country, with smaller groups found in Argentina, especially in Neuquén province.

A great many other Indian languages also are spoken by members of numerous smaller groups, many of which are extremely localized and some of which are on the verge of extinction. These groups are found primarily on the periphery of lowland regions, in areas once isolated from slave trading and the rubber trade. Relatively few lowland groups are located in Brazil, the rest being found in the Hispanic countries. Among the larger groups of the Amazon basin (excluding Quechua speakers) are the Chiquitanos of eastern Bolivia, the Arawaks (Campa, Machiguenga, etc.) and Shipobo of east-central Peru, the Cocama-Cocamilla of northeastern Peru, the Jívaroans along the Ecuador-Peru border, the Tikuna of the Brazil-Colombia-Peru border region, the Yanomamo of the Venezuela-Brazil border region, and the Makushí along the Brazil-Guyana border. Groups south of the Amazon basin include the Chiriguano of southeastern Bolivia and northwestern Argentina and the Toba of northern Argentina. North of the Amazon basin are the Arawaks of Guyana, the Goajiro and Sinu (Cenú) of northern Colombia, and the Emberá along the Colombia-Panama border. The Quillacingas of Colombia occupy lands just to the north of the Quechua domain.

Shamanism

Religious patterns. South American Indians traditionally practiced shamanism, a belief system in which chosen individuals attempt to use esoteric knowledge to cure illness and to avert harm. Indigenous cultures also practiced rites of passage and seasonal rites, animal (and occasionally human) sacrifice, and ceremonial drinking. The Inca built temples, maintained a priesthood and a class called Chosen Women who were dedicated to the service of the gods, and performed pilgrimages.

The Iberian invaders considered most Indian practices to be superstition rather than heresy, and over time they were able to convert the indigenous population to at least an external observance of Roman Catholicism. Perhaps 85 percent of the South American population now professes Catholicism. In Brazil, tens of millions of people combine Catholicism with African elements in such cults as Macumba and Candomblé. New movements such as liberation theology have enhanced the popularity of Catholicism in many communities. Protestantism long has been present in the Guianas, and it has become more widespread in the late 20th century, especially in Chile, southern and southeastern Brazil, and among the Aymara of Bolivia and Quechua speakers of central Ecuador. The largest Jewish community is in Argentina in Buenos Aires.

Sociological changes. All or most legal discrimination against the Indians and other ethnic sectors of the population was nominally abolished either at the time when the individual countries became independent or during the 19th century. The real conditions of the Indians (and to a certain extent of the Africans after the abolition of slavery in Brazil in 1888), however, remained the same or became worse, since, on one hand, liberal legislation tended to eliminate all communal property and the legal existence of the Indian communities, while, on the other, various forms of exploitation continued unchanged. These de facto conditions also were reinforced by 19th-century pseudo-scientific doctrines, one of which claimed that Indians and Africans were biologically inferior races.

In the 20th century, however, a partial change in intellectual attitudes and political conditions has resulted in initiatives toward ameliorating the conditions of these groups. In Brazil, for example, institutions such as the Protective Service for the Indians (Serviço de Proteção do Indio) and the National Indian Foundation (Fundação Nacional do Indio) have been established, although such organizations often have become agents for the relocation and control of Indian groups rather than for their interests and survival. Christian missionaries sometimes have acted as representatives of Indian rights. Indians of the Andean highlands have benefited from land reforms enacted since 1950 in Bolivia, Peru, and Ecuador, although these reforms often have defined rural peoples as "peasants" rather than as Indians. Such groups as the Agaruna and the Shipibo in eastern Peru have been able to take advantage of programs by which some Indians actually have become the landlords of mestizos. National parks and protected areas have been established for such peoples as the Yanomamo of Brazil and Venezuela and the Huaorani of Ecuador.

Indian-rights movements

Large-scale Indian-rights movements have appeared in the highlands, which have attempted to link different Quechua-speaking groups into broader unions in order to obtain land and political recognition; often these movements have claimed that Indian groups constitute nations in their own right. Lowland Indians also have organized—as in the Kayapó of Brazil and the Shuar of Ecuador—and larger, pan-Indian movements have emerged that have striven to unite disparate groups at the national and international level. Coupled with the rise of these movements has been a growing interest in Indian languages, technology, music, and medicine and an effort to use indigenous knowledge to provide appropriate economic development and help conserve the South American landscape.

(G.Ge./Gr.W.K.)

DEMOGRAPHIC PATTERNS

The continent's demographics reflect an unusual settlement history: South America is a "hollow continent," with most of the population concentrated around its margins. The highest population densities are found in the old Indian core areas of the Andes, the former slave areas of northeastern Brazil, and the areas of European immigration in southern Brazil, Uruguay, and Argentina. The interior is relatively empty because of the decline in Indian populations, poor communications with coastal areas, and the absence of economic opportunities capable of attracting large numbers of immigrants. Another characteristic of South American demography is a high rate of population growth in tropical regions coupled with moderate growth in the temperate southern cone. The high tropical growth rates, however, have begun to diminish.

Both South American demography and history can be explained through the changing patterns of birth and death rates and immigration caused by the Iberian conquest and by subsequent economic development. After the conquest, diseases such as smallpox, measles, malaria, and yellow

fever decimated all Indian populations, leading to a long-term pattern of high death rates and declining or stagnant populations, even where fertility was high. Beginning in the area of European migration and extending throughout the continent after World War II, innovations in public health, such as safe drinking water and vaccines, have resulted in a dramatic drop in death rates everywhere except remote rural areas and urban slums.

The demographic transition and fertility. The resultant population explosion has been caused by a traditionally high fertility rate and a modern low mortality rate. The situation can be thought of as the second of three stages. The first stage, which characterized most of South America during the 18th and 19th centuries, involved a rough population balance maintained by high death and birth rates. The second, transitional, phase has consisted of a population explosion brought about by declining mortality and continued high fertility. The third constitutes a modern stage where low fertility brings population stability.

Fertility rates are a result of many factors, including the availability and cultural acceptability of birth-control measures. In general, economic factors increasingly have come to dominate in decisions regarding family size, including the benefits of young children to families and the costs of rearing children to adulthood. Children traditionally have had considerable value in helping families in their farming or urban-craft livelihoods and in providing security for the elderly. The costs of maintaining children in such circumstances were low, and women had few opportunities outside the household to compete with child-raising.

Changes in fertility in South America have occurred with the expansion of mandatory education, which simultaneously has raised the cost of rearing children, reduced their benefits to families, and provided young women with the education needed to seek employment. A decline in birth rates has occurred despite opposition to birth control on the part of the Roman Catholic church. Birth rates declined first in the more prosperous areas of the southern cone, but all regions have experienced the effects of growing educational and employment opportunities for young women in an increasingly urbanized environment.

Effects of rapid population increase. Rapid population increase has had important demographic and social effects. Two examples are especially illuminating.

At the peak of population growth during the second stage, the proportion of children tends to be high, while in the third stage it is low. In South America the proportion of the population under 15 years is relatively high. As a consequence, the group of people in their productive (working) years is greatly reduced. This high ratio creates

Population density of South America.

a heavier burden for the working group, while the economy is not able to raise the productivity level needed to compensate for it.

Urban growth

Another crucial consequence is the so-called urban explosion. Argentina and Uruguay have become two of the most urbanized countries in the world, but their urban growth has been the result of mass foreign immigration. The dramatic increase in urban concentration began approximately in the 1930s. In all of Latin America, the proportion of urban centres with more than 10,000 inhabitants increased from one-fourth in 1950 to about three-fourths in 40 years.

South America now is one of the most urbanized regions in the world, following the industrially advanced areas. Although the rate of growth in larger cities has decreased since 1950, the urban population has continued to be concentrated in the larger cities: a large proportion of the urban, and in some cases of the total, population lives in a single urban centre. This situation prevails in Uruguay, Argentina, Chile, Paraguay, and Peru.

Much of the growth of South America's largest cities has come about from natural increase. Migration has remained important, however, because of unfavourable conditions in the countryside. Land ownership often has been concentrated in a few hands, while "urban bias" in national policy has been reflected in price controls for or subsidized imports of foodstuffs, concentration of health, recreational, and educational facilities, and the expansion of government bureaucracies in large cities. Attempts to reverse urban bias have met with strong political opposition. There also have been pressures from rural guerrilla movements, especially in Peru, which has caused an especially rapid migration to Lima.

The unequal distribution of population in South America—the "hollow continent" phenomenon—is likely to continue and even become more pronounced. Although certain frontier areas, such as Rondônia state in Brazil and the coca-growing regions of the Andes, have attracted substantial in-migration, these flows have been far less than the out-migrations to towns and cities in already densely populated areas. Development of the interior increasingly has relied on labour-saving technology, resulting in little incentive for migration. As the largest cities face mounting problems, the likely solution will be the urbanization of nearby rural centres. (Gr.W.K.)

The economy

As a group, the economies of South American nations have changed profoundly since the 1970s. This has come as a result both of external conditions beyond the control of these nations and of internal policy decisions made to produce change. At the most fundamental level, these countries mainly are exporters of relatively low-value primary products and semiprocessed materials and importers of higher-value manufactured goods. Great efforts have been made across the continent to expand the manufacturing sectors and to reduce dependency on imports.

From the 1930s until the late 1980s, the majority of South American nations pursued economic development strategies based on a system of import substitution. National governments used such measures as tariff and price policies to encourage domestic industries and protect them from external competition. They also created joint ventures with private capital and established state-owned enterprises, especially in the heavy industries, utilities, and transportation. They provided high subsidies for social programs in areas such as education and public housing. Furthermore, national spending on armaments and "defense" soared during periods of military rule.

Nations borrowed from foreign private banks and international lending institutions, such as the World Bank and the Inter-American Development Bank, to fund existing programs while also trying to expand their economic productivity through investments in areas such as transportation, energy generation, industrialization, and agricultural modernization. However, many countries lived well beyond their means through the wholesale borrowing of funds at high interest rates on the world market. Consequently, they were forced to borrow more and more money just to service the interest payments that accumulated annually on their outstanding debt, thus creating the so-called "debt crisis."

The debt crisis

With the debt crisis, economic chaos befell many South American nations. After decades of substantial progress in its economic development, the region as a whole regressed significantly in the 1980s. Between 1980 and 1990, gross domestic product (GDP) per capita measured in constant dollars declined for every South American nation except Brazil, Colombia, and Chile. For a part of this same period, inflation rates skyrocketed in many nations, exceeding 3,000 percent per year in some instances. Currency devaluation, economic austerity programs, and governmental disinvestment were the most commonly used remedies to check these problems.

The severity of their problems and the demands of their lenders have prompted most South American countries to initiate fundamental restructurings of their economies. These reorganizations are in accord with neoliberal, or "free-market," economic theory, which has come to dominate the region's economic planning and decision making in the 1990s. Emphasis has been placed on stimulating economic growth while selling state-owned enterprises to private investors and eliminating or severely curtailing support for social programs. The long-term goal is to increase productivity, reduce governmental expenditures, and diversify economic activities. Regional economic integration also has taken on new importance. The impact of these changes is not yet clear, but their most dramatic effects are felt by those lowest on the socioeconomic ladder. (E.C.G.)

Toward a free-market economy

RESOURCES

Mineral resources. South America is relatively rich in mineral resources. However, they are highly localized: few nations have a good balance of fuels and raw materials within their boundaries, and two countries, Uruguay and Paraguay, are nearly devoid of mineral wealth.

Mineral fuels. Large quantities of oil and natural gas are found in several areas within South America. The greatest quantities are located in the sedimentary materials surrounding Venezuela's Lake Maracaibo and the adjacent Caribbean coastal margin. Venezuela also has major deposits of oil and natural gas in the area surrounding El Tigre. Venezuela is one of the world's largest oil producers and exporters. Since 1972 Ecuador has also become a major oil exporter, utilizing newly discovered fields in the Amazonian region east of the Andes. Oil fields were brought into production at the same time in the Peruvian portion of the Amazon basin west of Iquitos. Bolivia produces oil and natural gas in its eastern lowlands around Santa Cruz, while Argentina and Chile share significant deposits bordering the Strait of Magellan in Patagonia and Tierra del Fuego. Additionally, Argentina has traditional oil-producing regions around the Patagonian city of Comodoro Rivadavia. Brazil has limited offshore oil and gas reserves. Colombia has long been self-sufficient in oil and gas production, with primary areas in the central Magdalena River valley and the Putumayo area adjacent to its border with Ecuador.

South America is poor in coal. Colombia exports coal from La Guajira Peninsula and the lower Magdalena River basin south of Barranquilla, and Argentina has limited quantities of good-quality coal at El Turbio in the extreme south. Brazil produces relatively small quantities of coal in its southern states, while areas in northwestern Venezuela and inland from Concepción in Chile also have coal mines.

Iron and ferroalloys. South America contains about one-fifth of the world's iron ore reserves. The most important beds are located in Brazil and Venezuela, supplying domestic iron and steel industries as well as significant exports. The great majority of the continent's reserves are in the Brazilian states of Minas Gerais, Pará, and Mato Grosso do Sul, where lodes of magnetite and hematite ores contain up to 50 to 65 percent iron content. In Venezuela, sites like Mount Bolívar and El Pao in the Sierra de Imataca at the base of the Guiana Highlands have re-

Iron ore deposits

Basic structural regions and principal mineral and hydroelectric sites of South America.

serves of ore containing a high percentage of iron. High-quality beds of this type are also found at Mutún, Bol., and in the central part of the Chilean Andes. Oolitic iron ore (*i.e.,* ores consisting of small round grains cemented together) is found at Sierra Grande in Argentina and Paz del Río in the Cordillera Oriental of Colombia. In addition, important iron ore deposits are located at Marcona, Peru, and along a narrow belt from Taltal to Ovalle in northern Chile. Lateritic deposits of ferrous hydroxides are widespread, mainly in Colombia, Brazil, and Argentina.

Among ferroalloys, manganese occurs in sedimentary forms in the Brazilian states of Amapá and Minas Gerais, as well as in highland Bolivia. It is also found in much lesser quantities in Argentina, Chile, Ecuador, and Uruguay. Chile has the second largest reserves of molybdenum in the Americas, trailing the United States.

Copper reserves

Nonferrous base metals. Nonferrous metals are abundant in South America. The continent's copper reserves represent more than one-quarter of the world's known reserves, nearly all of which are found in Chile and Peru. In Chile the Chuquicamata deposits of the northern Atacama Desert contain the largest amounts of copper known in the world and have ores containing 2.5 percent copper. Peru's most important deposits are found in the country's central Andean ranges near Yauricocha, Morococha, Casapalca, Cerro de Pasco, and Huarón, as well as in the south, where the Toquepala mine opened in 1960. Copper also is found in Argentina and highland Bolivia and at Caraíba in the Brazilian state of Bahia.

Bolivia ranks among the world's four or five largest tin producers. Sedimentary deposits and veins of tin occur in the country's eastern Andes in a narrow belt extending roughly from Oruro to Potosí. Significant tin deposits also occur in Brazil's western Amazon basin near the Madeira River. Lead and zinc are dispersed among many countries but are found in greatest abundance in the central Andes of Peru; in the state of Minas Gerais, Brazil; in highland Bolivia; and in the northern Argentine Andes. Bauxite is exploited in Guyana and Suriname; there are also major production centres near Cuidad Guayana in Venezuela and in several places in the eastern Amazon basin of Brazil. Small quantities of mercury occur in the central Andes of Chile and near Huancavelica in Peru. Antimony is found in Bolivia and in lesser quantities in central Peru.

Gold and silver

Precious metals and gemstones. Despite the fact that South America was Europe's treasure trove for gold and silver from the 1530s through the late 1700s, in the late 20th century the region contributes only a small percentage to the world's production of these precious metals. Brazil is South America's leading gold producer, with deposits in the Amazon basin accounting for much of the output. Traditional mining centres in Minas Gerais,

Goiás, and Mato Grosso have diminished in importance. Placer deposits in Colombia's Atrato River basin are significant sources of gold, and the metal is still produced also in Venezuela and in classical gold-mining centres in the central Andes of Peru, in the Andes of Chile, and in the Carajás area in Brazil. Peru has historically been one of the world's main silver exporters, primarily from mines extending from Cerro de Pasco to Huancayo in the Andes, but production has decreased since the early 1970s. Ecuadorean silver is located primarily in the Andes, while Colombia, Argentina, and Bolivia also exploit this metal in their highland areas. Platinum is found in the Cordillera Occidental of Colombia as well as in smaller quantities in association with the copper-mining activities of central Peru.

Many parts of South America, mainly in Brazil, are famous for their gems. The ancient bedrocks of the Brazilian Highlands, especially in the states of Minas Gerais and Goiás, are rich in precious and semiprecious stones, including diamonds. However, Brazil contributes only a small percentage to world diamond production.

Other precious or semiprecious stones abound in the same region, notably topazes, tourmalines, beryls, aquamarines, chrysoberyls, garnets, opals, and sapphires, as well as quartz of sufficiently high grade for use in the electronics industry. Colombia is famous for its emeralds, found primarily in the Muzo mines of the Cordillera Oriental.

Various elements used in industry, such as beryllium, niobium (columbium), tantalum, thorium, lithium, rare-earth metals, and mica, are extracted in South America. Brazil, from the Northeast through Minas Gerais, and Argentina's Sierra de Córdoba are important sources for these minerals.

Nonmetallic deposits. South America has few significant reserves of phosphates or potash, the primary bases for many fertilizers. The Atacama Desert of northern Chile has large deposits of nitrates, which were used for fertilizers and in the munitions industry until they were largely replaced by synthetic fertilizers and atmospheric nitrogen; however, they continue to be used as a source of iodine. Limestone is quarried in Ecuador, Brazil, Chile, and Colombia, while Uruguay and Colombia are significant producers of marble. Brazil has the continent's only significant deposits of graphite.

Biological resources. South America's abundant and diversified biological reserves have been described previously. These resources, however, are unevenly distributed throughout the continent; for example, the only large areas suited to wide-scale agriculture are found in the Argentine Pampas, central Chile, southeastern Brazil, and the littoral of Uruguay.

Botanical resources. Pre-Columbian cultures domesticated numerous plants, such as corn (maize), potatoes, cassava, and beans, which, when introduced into the Old World, became dietary staples there and revolutionized the world's food supplies. In addition, a great number of South American plants provide valuable drugs, including quinine (obtained from the bark of several trees of the genus *Cinchona,* indigenous to the eastern slopes of the Andes) and cocaine (extracted from the leaves of the coca shrub, found in the eastern Andes from Peru to Bolivia).

Domesticated plants of pre-Columbian cultures

The extensive forests that cover about half of the continent constitute South America's richest natural resource. With more than 1.5 million square miles of tropical rain forest, Brazil is the most densely forested country in the region. However, since the 1980s rapid deforestation in the Amazonian rain forest has become a worldwide concern because of its effects on the environment, including the loss of biodiversity and potential climatic change. Softwood forests, though much more limited, are extensive south of the Bío-Bío River in Chile and in the southern Argentine Andes, as well as from Paraná southward in Brazil. Tropical grasslands, such as the savannas of the Llanos of Colombia and Venezuela, Brazil's Mato Grosso Plateau, and the Argentine Pampas, a midlatitude grassland, represent South America's second major botanical resource.

Animal resources. Animal resources constitute a major portion of the economies of most South American countries. In pre-Columbian times, relatively few animals were domesticated, and almost none of them extended beyond the geographic limits of their wild ancestors. An exception was the Muscovy duck. Llamas and alpacas were domesticated in the high Andes in Inca times and probably earlier. Guinea pigs were domesticated as a meat source in pre-Inca times in the Andean highlands from Colombia to Argentina. Among animals introduced to the continent were cattle, horses, goats, sheep, and pigs, all of which adapted rapidly and thrived in the New World. Cattle have become especially important in areas such as the Llanos in Venezuela and Colombia, the Argentine Pampas, and the rolling plains of Uruguay, while sheep and goats predominate on the drier, colder grazing lands of the high mountainous areas and Patagonia. Game is plentiful in most habitats, though mammals are few and specialized. Deer are represented by several species. A variety of animals are exported as pets or for zoos.

Several Neotropical animals provide world-famous fur or wool. Chinchilla, native to the high Andes from Peru to northern Argentina, were hunted for their delicate gray fur to the point of near extinction. Vicuñas continue to be hunted despite protective laws and a ban placed on

Large iron mine in the Serra dos Carajás, Pará state, Brazil.

the trade of their fur. Efforts are being undertaken to increase their numbers by "ranching" vicuñas. The giant otter of the Amazon, several spotted cats such as the ocelot and jaguar, and rodents like the nutria also provide highly prized furs.

Aquatic animals

The colonies of seabirds along the Peruvian and Chilean coasts produce an accumulation of dung (guano) which is an important fertilizer. Pinnipeds (a suborder of aquatic carnivorous animals, including seals and walruses) are exploited for their oil and furs, particularly in Uruguay. Fur seals and sea lions are found along the southern coasts of South America, although their numbers have been reduced by hunting. (E.C.G./J.P.D.)

FORESTRY, FISHING, AND MINING

Although the Neotropical forests are renowned for their biological diversity, the bulk of their trees consists of fewer than 200 species. The mixed character of these forests is a major obstacle to large-scale exploitation of timber. Nonetheless, timber harvesting has expanded dramatically since 1950, especially in the Amazon basin. Many species are used as cabinet woods, including the highly prized mahogany from Venezuela, Brazil, Peru, and Bolivia and several leguminous species such as rosewood. Some species are exploited as general utility woods and are mainly used domestically, often as fuel. Other species, such as the quebracho tree found in the Gran Chaco of Argentina and Paraguay, which produces tannin, have significant commercial value. Commercial tree plantations have become important sources of forest products, especially in Argentina and Chile. Additionally, eucalyptus groves have been planted throughout the region since their introduction in the early 1800s and provide both building material and fuel.

Commercial tree plantations

Freshwater fish, abundant in many South American rivers, have been exploited as a food source since the earliest times, especially in the Amazon region and in the Guianas. Trout were introduced by Europeans into Andean lakes and rivers, sometimes to the detriment of endemic species, while reservoirs in northeastern Brazil and elsewhere have been stocked with tilapia from Africa. Most freshwater fishing is for local consumption.

Marine fisheries became important in the 1960s, when Peru emerged as one of the world's major fishing nations, based on its anchovy fisheries. However, overexploitation has severely depleted this resource. Chile has developed a large commercial marine fishing industry as well as salmon, trout, and shrimp "farms" aimed at the export market. Since 1980 Ecuador has been a leader in shrimp exports.

Mining, including the extraction of hydrocarbons, constitutes a significant portion of the gross national product of several countries, most notably Venezuela and Chile. Oil and natural gas are the most valuable minerals recovered. Venezuela is the region's largest petroleum producer, while Brazil leads in iron ore tonnage, accounting for more than four-fifths of the output. Brazil also is the largest producer of bauxite and tin concentrates, although Bolivia's tin output represents a large share of South America's total. Chile produces approximately three-fourths of the continent's copper output.

AGRICULTURE

Agriculture constitutes a large portion of South America's economy. Livestock production also occupies large parts of rural South America, especially cattle ranching. Most of the commercial livestock production, especially for the export sector, occurs on huge estates that have been the source of economic and social dominance for their owners for many generations.

Only about one-eighth of South America's land is suitable for permanent cropping or grazing. It is broadly agreed that agricultural land use throughout the continent is less efficient than it might be. Farm and ranch productivity could be enhanced by measures such as providing adequate agricultural credit, improving marketing, storage, and transportation systems, and expanding the educational system in rural areas. Such changes would benefit the large number of small farmholdings (*minifundias*)—three-fourths of South America's farmers own less than 25 acres (10 hectares)—making it possible for those farmers to improve their living standards and contribute to national development. The changes also would help to alleviate the widespread under- and unemployment prevalent in the densely populated rural areas. Unemployment is a problem in such areas, even though less than one-third of South America's working population is employed in the agricultural sector, as compared with nearly one-half of the population for the world as a whole.

Inefficient land use

The agricultural sector is affected negatively as well by the unfavourable terms of trade between agricultural commodities and manufactured goods that have existed in general since World War II. The rise in the cost of farming has outstripped the rise in the prices paid for agricultural commodities, and this imbalance substantially lowers the investment potential in the agricultural sector.

Principal crops. *Food crops.* Corn (maize), a native of tropical America and now a staple in countries around the world, is the most widely cultivated crop throughout the continent. Argentina became a major exporter of corn during the 20th century. Beans, including several species of the genus *Phaseolus,* are widely cultivated by small-scale methods and form an important food item in most countries. Cassava and sweet potato also are indigenous to the New World and have become the basic foodstuffs of much of tropical Africa and parts of Asia. The potato, which originated in the high Andes, became a dietary staple of many European nations. Several other plants were domesticated in South American environments, such as quinoa and canahua, both small grains used as cereals, and tuberoses such as ullucu and oca. Squashes and pumpkins are pre-Columbian crops that have spread throughout the world, as is the tomato, indigenous to South America's west coast. Cashews, cultivated in most tropical countries, and Brazil nuts, harvested from trees in the Amazon basin, are widely regarded as delicacies, but both the cashew fruit and the nuts also are local favourites. Cacao, native to the Amazon region and the source of cocoa, was prized by indigenous peoples and is still cultivated in many parts of South America, particularly in the state of Bahia, Brazil. Avocados also originated in the same region. Pineapples, probably indigenous to southern Brazil and the Paraná River basin, were cultivated throughout tropical South America and the West Indies prior to the arrival of Columbus. Papaya and guava are also from tropical America.

Introduced species

Europeans introduced a number of plants to the continent. Sugarcane has been cultivated in the humid tropics of South America since early colonial times, especially in northern Brazil, where it became the mainstay of the economy. In similar environments bananas have long been an important local food item, and since the early 1970s Ecuador has become one of the largest banana exporters in the world. Mangoes, oranges, lemons, and grapefruits are grown widely throughout tropical and subtropical environments in South America. While their origin is much disputed, coconuts are common in most tropical coastal areas in the region. Among the cereals, rice, which was introduced from Asia, has become a dietary staple in several countries. It is grown extensively in the irrigated desert oases of the Peruvian coast and in savanna and rain forest climatic areas of Brazil and Colombia. Wheat, along with other cereals, was introduced by the Spanish by the 1550s throughout Andean South America, where it is still grown. However, it is most successfully produced on the Pampas of Argentina and the littoral of southwestern Uruguay, where it was introduced after the mid-1800s. Soybeans were introduced in the Argentine Pampas in the 1950s and in southern Brazil in the 1960s and are now an extremely important commercial crop.

Specialized cash crops. Coffee was imported from the Old World in the 1800s and grown in the highlands of Venezuela, Colombia, and Ecuador. It is exported in great quantities from the main producing areas of Colombia's Cordillera Central, the source of some of the world's highest-quality coffees, and from several Brazilian states, including Paraná and Minas Gerais. The most notable native beverage, yerba maté, is brewed from the leaves of a plant indigenous to the upper Paraná basin. It is still

Agricultural regions of South America.

gathered in its wild state in Paraguay, Brazil, and Argentina, as well as grown on plantations in the latter two countries. Tobacco is cultivated in many countries but is produced commercially mainly in Brazil and Colombia. The two commercially most important native South American spices—allspice and red, or chili, pepper—are exported from Brazil.

South America also has a great variety of oil-producing plants, such as the babassu palm, native to Brazil, whose nuts are used for making soap. Vegetable waxes are produced mainly from the waxy secretions found on the leaves of the carnauba palm of Brazil. Vegetable ivory is taken from the hard seeds of the tagua palm, found in much of northern South America but particularly in lowland Ecuador.

Cotton, which has been used for cloth since prehistoric times, is grown in large quantities in northeastern Brazil, coastal Peru, and in the Chaco province of Argentina. Kapok and plants producing stem or leaf fibres, such as sisal, are extensively grown. The iraca, a plant with the appearance of a stemless palm, is cultivated in southern Colombia and northern Ecuador. Its fibres, extracted from young leaves, are woven into the once-famous "Panama" hats.

Natural latexes

Several plants furnish latex, from which rubber is extracted. Para rubber (seringa) and related species native to the Amazon basin were known by Indian groups and formed the basis for the Brazilian "rubber boom" of the late 1800s. Balata yields a nonelastic rubber used in golf balls and baseballs. Chicle, a latex gum extracted from the sapodilla tree, is used in the preparation of chewing gum. Artificial rubber has greatly reduced the demand for many natural latexes.

Livestock. Because cattle were of enormous cultural and economic importance for the Hispanic colonial economies, South America has a significant percentage of the world's total cattle population. Hybridized cattle breeds of the highest quality, such as Herefords, Angus, and Charollais, are found on the rich midlatitude pasturelands of the Argentine Pampas and in Uruguay. Much of the Llanos of northern South America is given over to the grazing of Brahman (Zebu) crosses. The pastures of the Amazon basin, created in the latter part of the 20th century, consist of imported African tropical grasses, which support large herds of Brahman, Charollais, and other hybridized breeds as well as exotics such as Asian water buffalo. Stock raising also flourishes in eastern and southern Brazil and in the temperate zones of the Andes; it is pursued in nearly all environmentally suitable areas of every country.

In the higher regions of the Andes, generally above 10,000 feet, and in the colder or more arid lowland settings, cattle

give way to other grazing animals. Llamas and alpacas, along with sheep and goats, are found in the higher Andes from Ecuador through northern Argentina and Chile. Vicuñas are still found at very high elevations, usually above 14,000 feet, in Peru. Sheep predominate in Patagonia and Tierra del Fuego, while goats prevail in the arid Atacama Desert and in the drought polygon of northeastern Brazil. Pigs and smaller domestic animals like chickens are present in almost all rural areas of the continent. Within the former Inca settlement area of the Andes, extending from the far south of Colombia through northern Argentina, guinea pigs are still raised as a food source.

INDUSTRY

In most South American nations, the industrial sector has made only a limited contribution to the creation of new sources of employment. This fact, which is problematic especially in view of the rapid growth of the labour force, can be explained in part by the adoption of production techniques requiring a high ratio of capital to labour and in part by the sector's slow growth. In the 1990s about one-fourth of the continent's labour force was occupied in the industrial sector.

In the early 1990s the industrial sector generated more than one-third of the gross national product for South America as a whole. Of this total, about four-fifths represented manufacturing and the rest construction and public utilities. In the two most industrialized countries, Argentina and Brazil, the production of foodstuffs, beverages, and tobacco accounts for only about one-seventh of the total manufacturing output. The metallurgy and mechanical industries represent more than one-third of total output, while chemicals and petroleum refining contribute about one-fourth and textiles, footwear, and apparel about one-eighth. During the last quarter of the 20th century, South American industrial production made substantial gains, especially in the output of cement and steel (ingots, rolled, plates, and sheets), pig iron, automobiles, and household appliances. Brazil, with its manufacturing core centring on São Paulo, has emerged as the industrial giant of the continent, followed by Argentina, Venezuela, and Chile.

The construction industry in much of South America retains fairly traditional methods. Construction techniques are labour-intensive, quality is often low, and costs are relatively high. Yet the building of high-rise and mid-rise office towers, hotels, commercial structures, and condominiums during the last two decades has greatly altered the skylines of virtually every large city on the continent. Despite the existence of a huge housing shortage, residential construction has lagged significantly behind demand. This is due largely to reduced governmental ability to provide state-financed housing and to limited involvement of speculation capital in residential construction.

POWER AND IRRIGATION

The total annual generation of electricity in South America increased by 10 percent between 1987 and 1990, mainly through the construction of large-scale hydroelectric projects. Industry consumes almost two-fifths of the electric power generated in South America, while household consumption accounts for about another two-fifths. The rest goes to other uses or transmission losses.

Energy consumption

Private installed electrical capacity, which consists mainly of relatively high-cost thermoelectric plants, represents a low percentage of total installed capacity. The overwhelming trend has been toward increases in hydroelectric power owing to the growing importance of public power services and the corresponding decline of private power generation.

Argentina, Brazil, Paraguay, Uruguay, and Venezuela have initiated ambitious electrical installation programs that have taken advantage of some of the region's abundant hydroelectric resources. From the 1960s through the 1980s, a major thrust of international lending institutions, including the Inter-American Development Bank and the World Bank, was to support infrastructural improvements. As a result, by the late 20th century South America boasted several of the largest hydroelectric projects in the world, including the Itaipu Dam on the Paraguay-Brazil border and the Guri Dam in the Llanos of Venezuela. The regional nature of many major rivers has led to binational cooperation on many projects, including joint efforts between Argentina and Uruguay and between Brazil and Paraguay.

Some of the hydroelectric projects also have been designed for irrigation uses, but there are many instances in which the construction of distribution canals for irrigation has yet to be achieved. Only 8 percent of arable land and of land under permanent crops is irrigated. Brazil and Argentina have the largest amounts of irrigated acreages, whereas Suriname, Peru, Chile, and Guyana have the highest proportion of irrigation in relation to cultivable land. Because of the relative scarcity of farmland in those countries, large-scale irrigation is a basic necessity of the agricultural sector.

TRADE

In South America, most banks and financial institutions are large enterprises with branches in many cities and towns. In some countries a high proportion of these were government-owned until the late 1980s. Foreign-owned banks or joint-venture enterprises of local and foreign capitalization are common. Wholesale and particularly retail business enterprises, on the other hand, are mostly individual concerns and in many cases are family shops. Department stores or chain stores, uncommon in most South American countries until the early 1970s, have become an important part of the merchandising environment, especially in the larger cities. During the 1980s, modern managerial and marketing structures took hold in many countries—especially Brazil, Argentina, Chile, Uruguay, Colombia, and Venezuela—often giving a competitive edge in the marketplace to enterprises that adopted them.

Internal trade. Intraregional trade continues to be of relatively small importance in South America. Despite its rapid growth since World War II, intraregional trade in Latin America still accounts for only about one-tenth of total exports. A firm conviction prevails in South America that intensification of intraregional trade is a necessary condition for rapid economic growth because it would help to reduce the region's excessive dependence on foreign markets, diversify exports, and alleviate balance-of-payments problems.

South American trade with the rest of Latin America is concentrated in a few countries. Argentina, Brazil, and Venezuela account for more than half of the exports, and these same three countries also absorb about half of the imports from the rest of Latin America.

All the independent South American countries except Guyana and Suriname belong to the Latin American Integration Association (LAIA; formerly the Latin American Free Trade Association). Despite formidable obstacles, including unequal levels of development, inadequate infrastructural linkages, and enormous physical distances between countries, the LAIA has directed its efforts toward designing a common trade policy for member countries. It will gradually reduce import duties and other restrictions on imports from the rest of the world while arriving at agreements to compensate trade payments between member countries as well as making reciprocal credit arrangements between central banks. In addition, Bolivia, Colombia, Ecuador, Peru, and Venezuela formed the Andean Group for the purpose of reaching agreement on common trade problems, including external tariffs, reductions in tariffs applicable to subregional production, and coordination of policies toward foreign investment. Peru suspended its membership in 1992 but remained as an observer.

Latin American Integration Association

About three-fourths of the trade among LAIA members consists of basic commodities and about one-fifth of semimanufactured and manufactured goods. Among basic commodities the most important items are foodstuffs, beverages and tobacco, raw materials, and nonferrous metals. Among manufactured goods the main items of trade are chemicals, machinery, automobiles, and transport equipment.

All these efforts toward integration have been responsible in part for the rapid growth of the area's internal trade and for a large contribution toward a greater balance of

trade among these countries. An important component of this trade has been binational barter agreements. Trade surpluses and deficits between the nations have declined considerably.

External trade. External trade represents a key element in South America's economic growth. Essential imports, particularly capital and basic intermediate goods, are needed to accelerate the industrialization process. A major problem has been that exports and net external financing have not generated sufficient income to pay for those imports. Despite increases in trade, South America's share of world trade has remained small, primarily because the volume of trade between major industrialized countries has grown at an even faster rate.

Exports

South America's major exports, in terms of value, are mostly primary commodities, including foodstuffs and plant products, fuels, and raw materials. Within the first group the most important commodities are sugar, bananas, cocoa, coffee, tobacco, beef, corn, and wheat. Oil, natural gas, and petroleum products dominate the second group, while linseed oil, cotton, cattle hides, fish meal, wool, copper, tin, iron ore, lead, and zinc top the third group. South American manufactured goods are gaining access to world markets as well. Brazil has become a significant supplier of armaments worldwide as well as an exporter of, among other products, small aircraft and shoes. Several other nations, including Uruguay, Argentina, Colombia, and Chile, have also increased their nontraditional exports over the past 20 years. More than one-fourth of these exports are sent to the United States, another one-fifth to western Europe, and one-twentieth to South America. Since the 1970s the illicit movement of drugs—particularly cocaine—which is mainly conducted from Peru, Bolivia, and Colombia, has added enormously to the value of exports from the region.

Almost three-fourths of South America's imports consist of machinery, vehicles and parts, chemicals and pharmaceuticals, paper and paperboard, textile products, and other manufactures. About one-fourth of South America's total imports are from the United States, one-seventh come from western Europe, and another one-seventh originate in South America. In general, South America's foreign trade sector has been slow to diversify; it is heavily dependent on imports for domestic supplies of industrial goods and suffers from an imbalance in trade with the industrialized countries.

TRANSPORTATION

In an area the size of South America, an efficient system of transportation is necessary for the development of the hinterland, the expansion of national markets, and the integration of the different national economic systems. Unlike North America, South America still does not have an adequately integrated transportation network. Significant efforts have been made to improve both the connections within countries and the linkages between them.

Roads. South America has an extensive and rapidly expanding network of roads. In many countries, however, only a relatively small percentage of the roads are paved, and in the most remote areas they may be barely wide enough for two vehicles to pass easily. The remainder of the system consists of improved roads or simply of dirt roads.

Highway integration

In developing national segments of international highways, particular attention has been paid to road-integration projects. The Inter-American Development Bank and the World Bank were heavily engaged in some of these projects, as, for example, in the construction of the bridge links joining Paraguay and Argentina, Argentina and Uruguay, and Paraguay and Brazil (all these links were completed by the late 1970s). A road linking Venezuela and Brazil allows north-south movement through the Amazon basin. Brazil continues to have the largest network of roads belonging to the Pan-American Highway system, which extends throughout the Americas.

Because of the size of the continent and the immense variety of physical environments, an efficient road network is of utmost importance. Roads not only provide the primary passenger routes for the great majority of people but also offer the most cost-effective means of moving goods within countries. In all South American nations, truck transportation has taken an increasingly large share of the volume of goods carried by land. In addition to stimulating economic development, routes such as the Transamazonian Highway and the Marginal de la Selva Highway linking the countries on the east side of the Andes, constructed since the 1970s, also represent attempts to spur development. The results of this effort, however, have been mixed, because making land available for settlement has often caused considerable ecological damage to the tropical forests.

Railways. In most South American countries railways have lost their dominant position as the major mode of transportation, having been replaced by the integrated road networks that have developed rapidly since the 1960s. Moreover, rail transport is plagued by operational problems as well as by obsolete equipment. Almost all lines are single-tracked, which makes traffic slow and discourages passenger service. Many countries have two or more track gauges, which impedes the efficient integration of the rail system.

Until the 1980s, virtually all railways were owned by the state. Since then, governments, as part of their overall efforts to privatize their national economies, have divested themselves of a large percentage of publicly owned railroads. This has led to the elimination of a huge number of passenger routes as well as the reduction of much of the freight component.

Maritime transport. Sea transportation has long been a vital component of the transport systems of South American countries. The great majority of imports and exports to and from the continent moves by ship. South America has a number of outstanding natural harbours, such as Rio de Janeiro, Salvador, Montevideo, and Valparaíso, along with numerous improved ports and roadsteads, including Buenos Aires, Callao, and Barranquilla. Many of these port facilities had degenerated significantly by the 1960s, to the point that some of the region's largest ports were blacklisted by insurers and shipping companies. Since the early 1970s, many of these ports have undergone extensive renovation and modernization, including the installation of containerization facilities.

Ports

Several countries are making a determined effort to develop and enlarge their national merchant marines. This effort is meant partly to arrest earlier trends of having their trade carried by ships from outside the region and partly to promote regional integration and improve the national balance of payments.

Waterways. There are two inland waterway systems of international importance, the Paraguay-Uruguay basin (which includes territory in four countries) and the Amazon basin (six countries). Each has several thousand miles of navigable waterways. Furthermore, there are three other minor systems: the Magdalena in Colombia, the Orinoco in Venezuela, and the São Francisco in Brazil. The remaining rivers are unsuitable for navigation. There are drawbacks to using some inland waterways, including dry seasons, the direction of water flow, and difficult rapids. In general, the volume of traffic on the waterways of South America is relatively small, and the prospects for increasing it are limited.

Air transportation. Air transportation has developed rapidly since World War II, mainly because it avoids the forbidding geographic obstacles that hamper surface travel. The increase in air transportation is particularly significant with respect to passenger traffic but applies less to the handling of bulky freight.

Each country has its own system of internal air services, operated until the late 1980s chiefly by government-owned or by heavily subsidized private companies. While several governments still operate an international carrier, privatization in the airline industry has spread to internal carriers. All the South American capitals and most of the large cities are linked by direct air services to the major traffic centres of the United States and Europe. Domestic traffic links have expanded extensively since the late 1970s, when "short take-off" jets were introduced into service.

(E.C.G./H.F.A.)

SOUTH AMERICAN GEOGRAPHIC FEATURES OF SPECIAL INTEREST

Landforms

ANDES MOUNTAINS

The Andes mountain system consists of a vast series of extremely high plateaus surmounted by even higher peaks that form an unbroken rampart over a distance of some 5,500 miles (8,900 kilometres)—from the southern tip of South America to the continent's northernmost coast on the Caribbean. One of the Earth's great natural features, the Andes separate a narrow western coastal area from the rest of the continent, affecting deeply the conditions of life within the ranges themselves and in surrounding areas. The Andes contain the highest peaks in the Western Hemisphere. The highest of them is Mount Aconcagua (22,831 feet [6,959 metres]) on the border of Argentina and Chile.

The Andes are not a single line of formidable peaks but a succession of parallel and transverse mountain ranges, or cordilleras, and of intervening plateaus and depressions. Distinct eastern and western ranges—respectively named the Cordillera Oriental and the Cordillera Occidental—are characteristic of most of the system. The directional trend of both the cordilleras generally is north-south, but in several places the Cordillera Oriental bulges eastward to form either isolated peninsula-like ranges or such high intermontane plateau regions as the Altiplano (Spanish: "High Plateau"), occupying adjoining parts of Argentina, Chile, Bolivia, and Peru.

Some historians believe the name Andes comes from the Quechuan word *anti* ("east"); others suggest it is derived from the Quechuan *anta* ("copper"). It perhaps is more reasonable to ascribe it to the *anta* of the older Aymara language, which connotes copper colour generally.

Map coverage of the Andes Mountains can be found below in the following sections: for the southern ranges, see *Patagonia;* for the central ranges, see *Amazon River basin;* and for the northern ranges, see *Orinoco River basin.*

Physical features. There is no universal agreement about the major north-south subdivisions of the Andes system. For the purposes of this discussion, the system is divided into three broad categories. From south to north these are the Southern Andes, consisting of the Fuegian and Patagonian cordilleras; the Central Andes, including the Chilean and Peruvian cordilleras; and the Northern Andes, encompassing the Ecuadorian, Colombian, and Venezuelan (or Caribbean) cordilleras.

Three subdivisions

Geology. The Andean mountain system is the result of global plate-tectonic forces during the Cenozoic Era (the last 66.4 million years) that built upon earlier geologic activity. About 250 million years ago, the crustal plates constituting the Earth's landmass were joined together into the supercontinent Pangaea. The subsequent breakup of Pangaea and of its southern portion, Gondwana, dispersed these plates outward, where they began to take the form and position of the present-day continents. The collision (or convergence) of two of these plates—the continental South American Plate and the oceanic Nazca Plate—gave rise to the orogenic (mountain-building) activity that produced the Andes.

Many of the rocks comprising the present-day cordilleras are of great age. They began as sediments eroded from the Amazonia craton (or Brazilian shield)—the ancient granitic continental fragment that constitutes much of Brazil—and deposited between about 450 and 250 million years ago on the craton's western flank. The weight of these deposits forced a subsidence (downwarping) of the crust, and the resulting pressure and heat metamorphosed the deposits into more resistant rocks; thus, sandstone, siltstone, and limestone were transformed, respectively, into quartzite, shale, and marble.

Approximately 170 million years ago this complex geologic matrix began to be uplifted as the eastern edge of the Nazca Plate was forced under the western edge of the South American Plate (*i.e.,* the Nazca Plate was subducted), the result of the latter plate's westward movement in response to the opening of the Atlantic Ocean to the east. This subduction-uplift process was accompanied by the intrusion of considerable quantities of magma from the mantle, first in the form of a volcanic arc along the western edge of the South American Plate and later by the injection of hot solutions into surrounding continental rocks; the latter process created numerous dikes and veins containing concentrations of economically valuable minerals that later were to play a critical role in the human occupation of the Andes.

The intensity of this activity increased during the Cenozoic—notably between about 15 and 6 million years ago—and the present shape of the cordilleras emerged. The resultant mountain system exhibits an extraordinary vertical differential of more than 40,000 feet between the bottom of the Peru-Chile (Atacama) Trench off the Pacific coast of the continent and the peaks of the high mountains within a horizontal distance of less than 200 miles. The tectonic processes that created the Andes have continued to the present day. The system—part of the larger Circum-Pacific volcanic chain that often is called the Ring of Fire—remains volcanically active and is subject to devastating earthquakes.

Physiography of the Southern Andes. The Fuegian Andes begin on the mountainous Estados (Staten) Island, the easternmost point of the Tierra del Fuego archipelago, reaching an elevation of 3,700 feet. They run to the west through Grande Island, where the highest ridges—including Mounts Darwin, Valdivieso, and Sorondo—are all less than 7,900 feet high. The physiography of this southernmost subdivision of the Andes system is complicated by the presence of the independent Sierra de la Costa.

The Patagonian Andes

The Patagonian Andes rise north of the Strait of Magellan. Numerous transverse and longitudinal depressions and breaches cut this wild and rugged portion of the Andes, sometimes completely; many ranges are occupied by ice fields, glaciers, rivers, lakes, or fjords. The crests of the mountains exceed 10,000 feet (Mount Fitzroy reaching 11,073 feet) north to latitude 46° S but average only 6,500–8,400 feet from latitude 46° to 41° S, except for Mount Tronador (11,453 feet). North of Lake Aluminé (Argentina) the axis of the cordillera shifts to the east up to a zone of transition between latitude 37° and 35° S, where the geographic aspect and geomorphic structure change. This zone marks the most commonly accepted northern extent of the Patagonian Andes; there is some disagreement, however, about this limit, some placing it farther south, at the Gulf of Penas, (47° S) and others considering it to be to the north, around 30° S.

The line of permanent snow becomes higher in elevation with decreasing latitude in the Southern Andes: 2,300 feet in Tierra del Fuego, 5,000 feet at Osorno Volcano (41° S), and 12,000 feet at Domuyo Volcano (36°38′ S). A line of active volcanoes—including Yate, Corcovado, and Macá—occurs about 40° to 46° S; the southernmost of these, Mount Hudson of Chile, erupted in 1991. Enormous ice fields are located between Mount Fitzroy (called Mount Chaltel in Chile) and Lake Buenos Aires (Lake General Carrera in Chile) at both sides of Baker Fjord; the Viedma, Upsala, and other glaciers originate from these fields. Other notable features are the more than 50 lakes found south of 39° S. Those depressions that are free of water form fertile valleys called vegas, which are easily reached by low passes. Magnificent and impenetrable forests grow on both sides of these cordilleras, especially on the western slopes; these forests cover the mountains as high as the snow line, although at the higher altitudes toward the north and in Tierra del Fuego the vegetation is lower and less dense. Both Argentina and Chile have created national parks to preserve the area's natural beauty.

Physiography of the Central Andes. The Central Andes

The Patagonian Andes in the Torres del Paine National Park, near Puerto Natales, Chile.

begin at latitude 35° S, at a point where the cordillera undergoes a sharp change of character. Its width increases to about 50 miles, and it becomes arid and higher; the passes, too, are higher and more difficult to cross. Glaciers are rare and found only at high elevations. There is virtually no vegetation, lakes having disappeared north of 39° S and forests north of 37°. The main range serves as the boundary between Chile and Argentina and also is the drainage divide between rivers flowing to the Pacific and the Atlantic. The last of the southern series of volcanoes, Mount Tupungato (21,555 feet) is just east of Santiago, Chile. A line of lofty, snowcapped peaks rise between Tupungato and the mighty Mount Aconcagua. To the north of Aconcagua lies Mount Mercedario (22,211 feet), and between them are the high passes of Mount Espinacito (16,000 feet) and Mount Patos (12,825 feet). South of Anconcagua the passes include Pircas (16,960 feet), Bermejo (more than 10,000 feet), and Iglesia (13,400 feet). Farther north the passes are more numerous but higher. The peaks of Mounts Bonete, Ojos del Salado, and Pissis surpass 20,000 feet; the snow line also is much higher.

The peak of Tres Cruces (22,156 feet) at 27° S latitude marks the culmination of this part of the cordillera. To the north is found a transverse depression and the southern limit of the high plateau region called the Atacama Plateau in Argentina and Chile and the Altiplano in Bolivia and Peru. The cordillera grows wider as it advances into Bolivia and Peru, where the great plateau is bounded by two ranges: the Occidental and the Oriental.

The Altiplano

Northward, to latitude 18° S, the peaks of El Cóndor, Sierra Nevada, Llullaillaco, Galán, and Antofalla all exceed 19,000 feet. The two main ranges and several volcanic secondary chains enclose depressions called salars because of the deposits of salts they contain; in northwestern Argentina, the Sierra de Calalaste encompasses the large Antofalla Salt Flat. Volcanoes of this zone occur mostly on a northerly line along the Cordillera Occidental as far as Misti Volcano (latitude 16° S) in Peru.

The western slopes of the Cordillera Occidental descend gradually to the Atacama Desert along the coast. At about 18° S the trend of the Cordillera Occidental changes to a northwesterly direction. The Cordillera Oriental to the east, lower and built on a broad bed of lava, is cut and denuded by rivers with steep gradients, fed by heavy rainfall. It has two sections. The southern portion is 150 miles wide and—with the exception of Chorolque Peak in Bolivia (18,414 feet)—of relatively low elevation. The northern section in Bolivia, called Cordillera Real, is narrow, with higher peaks and glaciers; the most important peaks, at over 21,000 feet, are Mounts Illimani and Illampu.

At about latitude 22° S the Cordillera Oriental penetrates into Bolivia and describes a wide semicircle to the north and then to the northwest; to the west the Altiplano reaches its broadest extent. The Altiplano—500 miles long and 80 miles wide—is one of the largest interior basins of the world. Varying in elevation from 11,200 to 12,800 feet, it has no drainage outlet to the ocean. Roughly in the centre of the plateau is a great depression between the two cordilleras. Lake Titicaca, the highest navigable lake of the world (110 miles long), fills the northern part of the depression; the Desaguadero River flows south through the depression, draining Titicaca water into the smaller Lake Poopó.

As the Andes enter Peru, the Cordillera Occidental runs parallel to the coast, while the Cordillera Real from Bolivia ends in the rough mountain mass of the Vilcanota Knot at latitude 15° S. From this knot (*nudo*), two lofty and narrow chains emerge northward, the Cordilleras de Carabaya and Vilcanota, separated by a deep gorge; a third range, the Cordillera de Vilcabamba, appears to the west of these and northwest of the city of Cuzco. The three ranges are products of erosive action of rivers that have cut deep canyons between them. West of the Cordillera de Vilcabamba, the Apurímac River runs in one of the deepest canyons of the Western Hemisphere. The city of Cuzco lies in the valley west of the Cordillera de Vilcanota at an altitude of nearly 11,000 feet.

The Peruvian Andes traditionally have been described as three cordilleras, which come together at the Vilcanota, Pasco, and Loja (Ecuador) knots. The Pasco Knot is a large, high plateau. To the west it is bounded by the Cordillera Huarochirí, on the west slope of which the Rímac River rises in a cluster of lakes fed by glaciers and descends rapidly to the ocean (15,700 feet in 60 miles). Ticlio Pass, at an altitude of some 15,800 feet, is used by a railway. Many small lakes and ponds are found on the knots, with Lake Junín (about 20 miles long) being the largest.

The Pasco Knot

North of the Pasco Knot, three different ranges run along the plateau: the Cordilleras Occidental, Central, and Oriental. In the Cordillera Occidental, at latitude 10° S, the deep, narrow Huaylas Valley separates two ranges, Cordillera Blanca to the east and Cordillera Negra to the west; the Santa River runs between them and cuts Cordillera Negra to drain into the Pacific. Cordillera Blanca is a complex highland with permanently snowcapped peaks, some among the highest of the Andes (*e.g.*, Mount Huascarán, at 22,205 feet). The glaciers that rise there often are broken off by earthquakes and rush down the slopes, demolishing vegetation and settlements in their paths. Cordillera Negra, so named because it is not covered with snow, is lower.

The two ranges join together at latitude 9° S. The Marañón River, which runs northward between the Cordilleras Occidental and Central at about 6° S, changes its direction of flow to the northeast, penetrating into a region of nar-

row transverse water gaps (*pongos*) that cut the cordillera to reach the Amazon basin. These include Rentema (about one and one-fourth miles long and 200 feet wide), Mayo, Mayasito, and Huarcaya gaps and—the most important—Manseriche Gap, which is seven miles long.

Between the Cordilleras Central and Oriental, the Huallaga River runs in a deep gorge with few small valleys; it cuts the eastern cordillera at Aguirre Gap (latitude 6° S). The Cordillera Oriental ends in the Amazon basin at 5° S.

The permanent snow line reaches an altitude of 19,000 feet in Mount Chanchani (about latitude 16° S) and declines to about 15,000 feet in Cordillera Blanca and to 13,000 feet on Mount Huascarán. Permanent snow disappears north of 8° S, the puna grasslands end, and the so-called humid puna, or *jalca,* begins. Mountains become wider and smoother in appearance, while vegetation changes to heathland and trees. The altitude diminishes, and passes are much lower, as at Porculla Pass (7,000 feet) east of Piura.

Physiography of the Northern Andes. A rough and eroded high mass of mountains called the Loja Knot (4° S) in southern Ecuador marks the transition between the Peruvian cordilleras and the Ecuadorian Andes. The Ecuadorian system consists of a long, narrow plateau running from south to north bordered by two mountain chains containing numerous high volcanoes. To the west, in the geologically recent and relatively low Cordillera Occidental, stands a line of 19 volcanoes, 7 of them exceeding 15,000 feet in elevation. The eastern border is the higher and older Cordillera Central, capped by a line of 20 volcanoes; some of these, such as Chimborazu Volcano (20,702 feet), have permanent snowcaps.

The Ecuadorian Andes

The outpouring of lava from these volcanoes has divided the central plateau into 10 major basins that are strung in beadlike fashion between the two cordilleras. These basins and their adjacent slopes, which are intensively cultivated, contain roughly half of Ecuador's population.

A third cordillera has been identified in the eastern jungle of Ecuador and has been named the Cordillera Oriental. The range appears to be an ancient alluvial formation that has been divided by rivers and heavy rainfall into a number of mountain masses. Such masses as the cordilleras of Guacamayo, Galeras, and Lumbaquí are isolated or form irregular short chains and are covered by luxuriant forest. Altitudes do not exceed 7,900 feet, except at Cordilleras del Cóndor (13,000 feet) and Mount Pax (11,000 feet).

North of the boundary with Colombia is a group of high, snowcapped volcanoes (Azufral, Cumbal, Chiles) known as the Huaca Knot. Farther to the north is the great massif of the Pasto Mountains (latitude 1°–2° N), which is the most important Colombian physiographic complex and the source of many of the country's rivers.

Three distinct ranges, the Cordilleras Occidental, Central, and Oriental, run northward. The Cordillera Occidental, parallel to the coast and moderately high, reaches an elevation of nearly 13,000 feet at Mount Paramillo before descending in three smaller ranges into the lowlands of northern Colombia. The Cordillera Central is the highest (average altitude of almost 10,000 feet) but also the shortest range of Colombian Andes, stretching some 400 miles before its most northerly spurs disappear at about latitude 8° N. Most of the volcanoes of the zone are in this range, including Mounts Tolima (17,105 feet), Ruiz (17,717 feet), and Huila (18,865 feet). At about latitude 6° N, the range widens into a plateau on which Medellín is situated.

Between the Cordilleras Central and Occidental is a great depression, the Patía-Cauca valley, divided into three longitudinal plains. The southernmost is the narrow valley of the Patía River, the waters of which flow to the Pacific. The middle plain is the highest in elevation (8,200 feet) and constitutes the divide of the other two. The northern plain, the largest (15 miles wide and 125 miles long), is the valley of Cauca River, which drains northward to the Magdalena River.

The Cordillera Oriental trends slightly to the northeast and is the widest and the longest of the three. The average altitude is 7,900 to 8,900 feet. North of latitude 3° N the cordillera widens and after a small depression rises into the Sumapaz Uplands, which range in elevation from 10,000 to 13,000 feet. North of the Sumapaz Upland the range divides into two, enclosing a large plain 125 miles wide and 200 miles long, often interrupted by small transverse chains that form several upland basins called *sabanas* that contain about a third of Colombia's population. The city of Bogotá is on the largest and most populated of these *sabanas;* other important cities on *sabanas* are Chiquinquirá, Tunja, and Sogamoso. East of Honda (5° N) the cordillera divides into a series of abrupt parallel chains running to the north-northeast; among them the Sierra Nevada del Cocuy (18,022 feet) is high enough to have snowcapped peaks.

Farther north the central ranges of the Cordillera Central come to an end, but the flanking chains continue and diverge to the north and northeast. The westernmost of these chains is the Sierra de Ocaña, which on its northeastern side includes the Sierra de Perijá; the latter range forms a portion of the boundary between Colombia and Venezuela and extends as far north as latitude 11° N in La Guajira Peninsula. The eastern chain bends to the east and enters Venezuela as the Cordillera de Mérida. On the Caribbean coast just west of the Sierra de Perijá stands the isolated, triangular Santa Marta Massif, which rises abruptly from the coast to snowcapped peaks of 18,947 feet; geologically, however, the Santa Marta Massif is not part of the Andes.

The northern end of the Andes

The Venezuelan Andes are represented by the Cordillera de Mérida (280 miles long, 50 to 90 miles wide, and about 10,000 feet in elevation), which extends in a northeasterly direction to the city of Barquisimeto, where it ends. The cordillera is a great uplifted axis where erosion has uncovered granite and gneiss rocks but where the northwestern and southeastern flanks remain covered by sediments; it consists of numerous chains with snow-covered summits separated by longitudinal and transverse depressions—Sierras Tovar, Nevada, Santo Domingo, de la Culata, Trujillo, and others. The range forms the northwestern limit of the Orinoco River basin, beyond which water flows to the Caribbean. North of Barquisimeto, the Sierra Falcón and Cordillera del Litoral (called in Venezuela the Sistema Andino) do not belong to the Andes but rather to the Guiana system.

Soils. The complex interchange between climate, parent material, topography, and biology that determines soil types and their condition is deeply affected by altitude in the Andes. In general, Andean soils are relatively young and are subject to great erosion by water and winds because of the steep gradients of the land.

In the Fuegian and southern Patagonian Andes, the formation of soils is difficult; the actions of glaciers and of strong winds have left nearly bare rock in many places. Peat bogs, podzols, and meadow soils, all with thick horizons (layers) of humus, are found; drainage is poor. Volcanic soils that are rich in organic material and are well drained occur in the region of lakes. North of latitude 45° S, soils are formed directly on weathered rocks at higher elevations, and reddish brown soils with gravel and quartz are found in the lower zones; erosion is heavy.

North of 37° S the Atacama Desert is covered with heavily eroded desertic soils that are low in moisture and organic material and high in mineral salts. This soil type, with few differences, extends along the Cordillera Occidental to north of Peru.

From Bolivia to Colombia the soils of the plateau and the east side of the eastern cordilleras show characteristics closely related to altitude. In the Andean *páramo* embryonic soils black with organic material are found. At altitudes between 6,000 and 12,000 feet, red, brown, and chernozem soils occur on moderate slopes and on basin floors. In more poorly drained locations, soils with a permeable sandy horizon are relatively fertile; these soils are the most economically important in Bolivia, Peru, and Ecuador. The *sabana* soils of Colombia are gray-brown, with an impermeable claypan in certain levels.

At high elevations soils are thin and stony. On the east side of the eastern cordilleras, descending to the Amazon basin, thin, poorly developed humid soils are subject to considerable erosion. Intrazonal soils (those with weakly developed horizons) include humic clay and solonetz (dark

Intrazonal soils

alkaline soils) types found close to lakes and lagoons. Also included in this group are soils formed from volcanic ash in the Cordillera Occidental from Chile to Ecuador.

The azonal soils—alluvials (soils incompletely evolved and stratified without definite profile) and lithosols (shallow soils consisting of imperfectly weathered rock fragments)—occupy the major part of the Andean massif. In Colombia, sandy yellow-brown azonal soils on slopes and in gorges are the base of the large coffee plantations.

Climate. In general, temperature increases northward from Tierra del Fuego to the Equator, but such factors as altitude, proximity to the sea, the cold Peru (Humboldt) Current, rainfall, and topographic barriers to the wind contribute to a wide variety of climatic conditions. The hottest rain forests and deserts often are separated from tundralike puna by a few miles. There also is considerable climatic disparity between the external slopes (*i.e.*, those facing the Pacific or the Amazon basin) and the internal slopes of the cordilleras; the external slopes are under the influence of either the ocean or the Amazon basin. As mentioned above, the line of permanent snow varies greatly. It increases from 2,600 feet at the Strait of Magellan, to 20,000 feet at latitude 27° S, after which it begins descending again until it reaches 15,000 feet in the Colombian Andes.

Precipitation varies widely. South of latitude 38° S, annual precipitation exceeds 20 inches, whereas to the north it diminishes considerably and becomes markedly seasonal. Farther north—on the Altiplano of Bolivia, the Peruvian plateau, and in the valleys of Ecuador and the *sabanas* of Colombia—rainfall is moderate, though amounts are highly variable. It rains only in small amounts on the west side of the Peruvian Cordillera Occidental and somewhat more in Ecuador and Colombia. On the east (Amazonian) side of the Cordilleras Orientales, rainfall usually is seasonal and heavy.

Temperature varies greatly with altitude. In the Peruvian and Ecuadorian Andes, for example, the climate is tropical up to an altitude of 4,900 feet, becoming subtropical up to 8,200 feet; hot temperatures prevail during the day, and nights are mildly warm. Between 8,200 and 11,500 feet daytime temperatures are mild, but there are marked differences between night and day; this zone constitutes the most hospitable area of the Andes. From 11,500 to 14,800 feet it generally is cold—with great differences between day and night and between sunshine and shadow—and temperatures are below freezing at night. Between about 13,500 and 15,700 feet (the puna), the climate of the *páramo* is found, with constant subfreezing temperatures. Finally, above 15,700 feet, the climate of the peaks and high ridges is polar with extremely low temperatures and icy winds.

As in other mountainous areas of the world, a wide variety of microclimates (highly localized climatic conditions) exist because of the interplay of aspect, exposure to winds, latitude, length of day, and other factors. Peru, in particular, has one of the world's most complex arrays of habitats because of its numerous microclimates.

Dependence of life on altitude

Plant and animal life. The ability of plants and animals to live in the Andes depends largely on altitude, although the existence of plant communities also is determined by climate, availability of moisture, and soil and that of animal life by the abundance of food sources; the permanent snow line is the upper limit of habitation. Some plants and animals can live at any altitude, and others can live only at certain levels. Cats rarely live above 13,000 feet, whereas white-tailed mice usually do not stay lower than 13,000 feet and can live up to 17,000 feet. The camelids (llama, guanaco, alpaca, and vicuña) are animals primarily of the Altiplano (11,200 to 12,800 feet), although they can live well at lower altitudes. It is thought that the condor can fly up to 26,000 feet.

Probably the low barometric pressures of high altitudes are less important for vegetation, but altitude imposes a number of climatic variables—such as temperature, wind, radiation, and dryness—that determine what kinds of plants grow in different parts of the Andes. In general, the Andes can be divided into altitudinal bands, each with typical predominant vegetation and fauna; but latitude imposes differences between south and north, and proximity to the Pacific and to the Amazon basin is reflected in differences between the external and internal slopes of the Cordilleras Occidental and Oriental.

A zone at about latitude 35° S separates two different regions of the Andes. To the south, in the Patagonian Andes, the flora is austral (of southern aspect) instead of Andean. Magnificent mid-latitude rain forests of the conifer genus *Araucaria* and of oak, coigue (an evergreen used for thatching), chusquea, cypress, and larch occur.

Characteristics to the north are different. The Cordillera Occidental is extremely dry in the south, slightly humid (with moisture and scarce rainfall) in central and northern Peru, and humid with heavy or moderate rainfall in Ecuador and Colombia. Vegetation follows the climatic scheme: in the south it is poor and desertlike, though at higher altitudes steppe vegetation occurs. Animals include the guemul, puma, vizcacha, cuy (guinea pig), chinchilla, camelids, mice, and lizards; among the birds are the condor, partridge, parina, huallata, and coot. Agricultural potential is poor. The east side of the Cordilleras Orientales northward from Bolivia has lush vegetation, most of it tropical forest with a rich jungle fauna.

On the plateau (valleys, plains, ranges, and internal slopes of the cordilleras), life again is closely related to altitude. Tropical palms and eternal snows lie within a few miles of each other, where altitude may vary from 1,600 feet in deep gorges to more than 20,000 feet in peaks and ridges. Up to an elevation of 8,000 feet, vegetation reflects the dry tropical and subtropical climate, and agriculture is important: the great coffee industry of Colombia is located mainly in the warm valleys of this zone. Between 8,200 and 11,500 feet lies the most populated zone of the Andes; some of the major cities of the Andean countries are there, and the zone supports the main part of Andean agriculture. Temperatures vary from warm in the valleys to moderate low (down to 50° F [10° C]) on the plains, *sabanas,* and slopes, and there is seasonal rainfall and water from rivers. This zone also is suitable for livestock and poultry farming.

The *páramo*

Between 11,500 and 13,400 feet relief is usually rough and difficult for agriculture. In Colombia this zone is *páramo* and sub-*páramo,* with seasonal rainfall; in Ecuador rain is abundant; and in Peru and Bolivia the *páramo* has from moderate to scarce rainfall. From 13,400 to 15,700 feet (the puna), vegetation consists of plants that resist the cold temperature and nighttime freezing; above 16,000 feet, vegetation is almost absent.

The people. Human presence in the Andes is relatively recent; the oldest human remains to be found are only 10,000 to 12,000 years old, although habitation probably dates to much earlier times. The shortage of oxygen at high altitude, especially above 12,000 feet, is so physiologically demanding that it imposes deep adaptative changes even within the cells of the body. The highest altitude in the Andes at which people have resided permanently is 17,100 feet (shepherds in southern Peru) and, as temporary workers, 18,500 to 19,000 feet (Carrasco Mine, in the Atacama Desert, Chile).

From Patagonia to the southern limits of the Bolivian Altiplano, the Andes are sparsely populated; a few small groups of shepherds and farmers live on the lower slopes and vegas of the cordillera. Farther to the north, from Bolivia to Colombia, the largest population concentrations and most of the important cities of these countries are found in the Andes. In Peru and Bolivia, a significant proportion of the inhabitants live above 10,000 feet.

Roughly half the population of Bolivia are Aymara- and Quechua-speaking Indians; most of the remainder are Spanish-speaking mestizos (or mixed). In the Lake Titicaca district live remnants of the ancient Uru people. Population is distributed mainly between the high *páramos,* where, except for a seminomad population of shepherds, the principal occupation is mining, and the lower narrow valleys, where the people practice agriculture. In Peru, mining is the most important human activity above 11,500 feet, but the great majority of the Andean population is engaged in agriculture and raising sheep, goats, llamas, and alpacas; a growing proportion of people have become

Farmland in the Peruvian Andes near Cuzco, with the Cordillera de Vilcambamba in the background.

employed in industry and commerce. A group of Aymara-speaking Indians live in the south around Lake Titicaca, but the largest native population is Quechua-speaking; Quechua speakers constitute the great majority of the highland population.

The inhabitants of the Ecuadorian Andes are mainly Quechua speakers and mestizos; in the south there are small groups of Cañaris and, in the north, Salasacas. Agriculture is the main occupation; some Indian peoples engage in ceramics and weaving.

In Colombia the largest proportion of the population lives between 5,000 and 10,500 feet. Only a tiny fraction of the country's population is Indian, these groups living on the Altiplano of the Cordillera Oriental and in the Cordillera Central and the southern mountains. The zone of coffee plantations at about 3,000 to 6,500 feet is the most densely populated area.

The economy. *Agriculture and livestock.* Agriculture on the Andes is difficult, and crop yields are relatively poor. The water supply is inadequate, and a large part of the plateau region is dry or receives little and irregular seasonal rainfall. Temperatures of the high plains are too cold, and crops are subject to freezing. The terrain is rough, and soils are not well developed; and, where fertile valleys do occur, they are narrow and small.

Thus, a considerable amount of Andean agricultural production is for local consumption. Some products, however, have been grown in sufficient quantity to be exported, including coffee (especially from Colombia), tobacco, and cotton; in addition, large quantities of coca (the source of cocaine) have been exported from Colombia and Bolivia, despite efforts to curb production. The possibilities of increasing the amount of arable land area by irrigation are extremely limited.

Export crops

The natural pastures of the plateau regions are extensively used for cattle raising. Colombia exports cattle, and Peru has a large milk-canning and livestock industry. Sheep, goat, llama, and alpaca raising are widespread in Peru and Bolivia, with both countries exporting sheep and alpaca wool.

Mining. The mining industry of the Andes is one of the most important of the world. Mining is especially extensive in the south. The principal minerals are copper in Chile and Peru; tin in Bolivia; silver, lead, and zinc in Bolivia and Peru; gold in Peru, Ecuador, and Colombia; platinum and emeralds in Colombia; bismuth in Bolivia; vanadium in Peru; and coal and iron in Chile, Peru, and Colombia. Extensive deposits of petroleum are distributed along the entire eastern side of the Andes.

Transportation. The Andes always have been a formidable barrier for communication, with great effect on the economic and cultural development of the region. Production centres generally are far from seaports, and the mountainous character of the land makes the construction and maintenance of railways and roads difficult and expensive. A large network of pack trails are still in use between small communities and between farms and markets. Horses, donkeys, and mules are widely used; in Colombia the ox and in Peru and Bolivia the llama also are transport animals.

Most of the railways have been built to transport mining products, and the internal systems otherwise are little developed. There are two international railways between Chile and Argentina: the first connects Valparaíso and Buenos Aires, and the second, Antofagasta and Salta. La Paz, Bolivia, is connected with Buenos Aires, Antofagasta and Arica (Chile), and (via Lake Titicaca) Arequipa and Matarani (Peru). Peru has two important internal railways, one from Puno to Cuzco and the other from Lima to Cerro de Pasco and Huancavelica; the latter line is the highest in the world, crossing Ticlio Pass at an altitude of some 15,800 feet. The main rail line in Ecuador runs from Quito to Guayaquil, and in Colombia the main line connects Bogotá to the Caribbean coast.

Roads are more suitable for Andean agricultural regions, because the small and widely separated valleys make railway construction and operation too expensive. Since World War II, all countries along the Andean cordilleras have expanded their road networks both within and through the mountains, although only a small portion of these roads are paved. The Pan-American Highway winds through the mountains of Colombia and Ecuador, connecting the major cities; various routes in Peru, Bolivia, Argentina, and Chile also are included in the system.

Air transport has become particularly important in the Andes, where it has reduced the difficulties of overland communication. Air routes are especially well developed in Colombia and Peru.

Study and exploration. As mentioned above, the Andes have been populated for millennia. By the time of the Spanish conquest in the 1530s, the indigenous highland peoples had developed a thorough knowledge of the Andes and had built in them an extensive network of cities and connecting roads. Early Spanish exploration of the mountains consisted of plundering raids, although in the process most of the major modern Andean cities were founded.

The first systematic European study of the mountains came in the form of a series of surveys called the

Relaciones geográficas (1579–85), which in increasingly elaborate questionnaires recorded much geographic and economic information about Spain's overseas colonies. In 1735 an expedition led by the French naturalist Charles-Marie de La Condamine began to measure the arc of the meridian at the Equator in the Andes, and for several years this group surveyed the Ecuadorian ranges. An even more important series of investigations was conducted by the German naturalist and explorer Alexander von Humboldt, who arrived on the Venezuelan coast in 1799 and for five years made innumerable observations of Andean geology, climatology, and biology (particularly of altitude-based ecological zones).

Explorations of Humboldt

By the mid-19th century the now-independent Andean countries were conducting and sponsoring scientific exploration of the mountains. Among those active at that time were the British mountaineer Edward Whymper in Ecuador, the Peruvian Mariano Paz Soldan in Peru, and the Italian geographer Agostino Codazzi, who produced detailed maps of Colombia and Venezuela. Since the late 19th century much Andean research has been directed toward economic development, primarily mining operations and railway construction. (M.T.V./N.R.S.)

GRAN CHACO

The Gran Chaco is an immense lowland alluvial plain in interior south-central South America. The name is of Quechua origin, meaning "Hunting Land." Largely uninhabited, the Chaco is an arid subtropical region of low forests and savannas traversed by only two permanent rivers and practically unmarked by roads or rail lines. It is bounded on the west by the Andes mountain ranges and on the east by the Paraguay and Paraná rivers. The Chaco's northern and southern boundaries are not as precise: it generally is said to reach northward to the Izozog Swamps in eastern Bolivia and southward to about latitude 30° S, or roughly the Salado River in Argentina. Thus defined, the Gran Chaco extends some 450 miles (725 kilometres) from east to west and about 700 miles from north to south and covers about 280,000 square miles (725,000 square kilometres); of this total, slightly more than half lies within Argentina, a third in Paraguay, and the remainder in Bolivia. The two permanent rivers, the Pilcomayo and the Bermejo (Teuco) flow southeastward across the Chaco from their Andean headwaters to the Paraguay River and demarcate the three main regional divisions of the Chaco in Paraguay and Argentina: the Chaco Boreal north of the Pilcomayo, the Chaco Central between the two rivers, and the Chaco Austral south of the Bermejo; the portion of the Chaco in Bolivia commonly is called the Bolivian Chaco.

For map coverage of the Gran Chaco, see below *Río de la Plata system.*

Physical features. *Physiography.* The Gran Chaco is a vast geosynclinal basin formed by subsidence (or downwarping) of the area between the Andean cordilleras on the west and the Brazilian Highlands on the east as it filled with alluvial debris from these two features. Because of its alluvial character, the Gran Chaco is nearly stone-free and is composed of unconsolidated sandy and silty sediments that are up to 10,000 feet (3,050 metres) deep in some places. The only rock outcrops of consequence are a few isolated remnants in Paraguay along the Paraguay River and some sandstone mesas in northern Paraguay and southern Bolivia.

Drainage. All but the extreme northwestern sector of the Gran Chaco is drained by west-bank tributaries of the Paraguay and Paraná rivers. The Bermejo and the Pilcomayo, even though they manage to traverse the Chaco, remain typical of most Chaco streams. Their courses are marked by countless sloughs, oxbow lakes, braided channels, sandbars, and vast swamplands; and they sustain such high losses from flooding, seepage, and evaporation that only a meagre portion of their full flow ever reaches the parent stream. Most of the Chaco is so poorly drained that the shallow, irregular channels on the exceptionally level plain lead to rapid and extensive flooding during the rainy southern summers (October to March). At the peak of these floods, as much as 42,000 square miles, or about one-seventh, of the area of the Chaco may be inundated, although some of this is caused as much by improper drainage of the impermeable subsoils as by overflow of the streams. Saline water is common in both deep and shallow wells, and the location and maintenance of freshwater supplies generally is a matter of chance. The problem appears to be greatest in the Chaco Boreal, although it has been suggested that the situation is more like that of the remainder of the Chaco or like the Argentine Pampa, where groundwater problems are not now considered to be as severe as early settlers and explorers had postulated.

Summer flooding

Soils. Chaco soils range from sandy to heavy clay. Soils in the more humid east have more organic material and lateritic subsoils, whereas in the west the soils contain less surface organic material and have predominately calcareous subsoils. The local determining factor is drainage, whether a function of soil texture or of relative relief. Sometimes differences in elevation of less than three feet result in different soil types. Grasslands, or savannas, generally tend to be associated with sandier soils, bushlands with poorly drained clay soils, and the forestland with better-drained clay soils. In many cases, the high concentration of dissolved salts in the groundwater creates conditions in swampy sites that are intolerable to most plants, thus extending an arid appearance even into many areas where water is abundant.

Climate. With its considerable north-south extent, the Gran Chaco is subject to climates that vary from tropical in the north to warm-temperate in the south. Most of the region, however, is subtropical. Average temperatures vary from 60° to 85° F (16° to 29° C), with an average relative humidity between 50 and 75 percent. Great temperature contrasts exist, and the highest recorded temperatures for the continent occur in the Chaco. Average maximums are near 80° F (27° C), and absolute maximums may reach 116° F (47° C). The average minimum is about 57° F (14° C), although freezing winter temperatures can occur throughout the region.

The highest average annual rainfall—52 inches (1,320 millimetres)—is in the east, and precipitation gradually decreases to about 20 inches in the far west. Although the rainfall normally would be adequate for agriculture, roughly a third to half of the total comes in the hot summer. Evaporation losses sharply reduce the effective precipitation and give the Chaco an arid nature that is absent only in the permanent swamps and forests along the Paraguay River.

Although light breezes are common, outbreaks of cool polar air from the south, called pamperos in Argentina, bring thunderstorms and strong gusty winds that occasionally exceed 60 miles per hour. These air masses move northward into the Amazon basin (where they are called

Chip and Rosa Maria Peterson

Palm savanna in the Chaco Central, near Formosa, in Argentina.

friagems). The windiest season, however, is spring, during the transition from warm to hot weather. Dust storms may occur in the dry season.

Plant life. The vegetation of the Gran Chaco is intimately associated with the pattern of soils and reflects the same general east-west division. The eastern Chaco is noted for its parklike landscape of clustered trees and shrubs interspersed with tall, herbaceous savannas. To the west, a wide transition zone grades into the *espinal,* a dry forest of spiny, thorny shrubs and low trees. Chaco vegetation is adapted to grow under arid conditions and is highly varied and exceedingly complex. The climax vegetation is called *quebrachales,* and consists of vast, low hardwood forests where various species of quebracho tree are dominant and economically important as sources of tannin and lumber. These forests cover extensive areas away from the rivers; nearer the rivers they occupy the higher, better-drained sites, giving rise to a landscape in which the forests appear to be islands amid a sea of savanna grasses growing as high as a person on horseback. In the more arid western Chaco, thorn forests, the continuity of which is occasionally broken by palm groves, saline steppes, and savannas induced by fire or deforestation, are dominated by another quebracho tree that has a lower tannin content and is used most often for lumber. There is also a marked increase in the number and density of thorny species, among which the notorious vinal (*Prosopis ruscifolia*) was declared a national plague in Argentina because its thorns, up to a foot in length, created a livestock hazard in the agricultural lands it was invading.

Sources of tannin and lumber

Animal life. True to its name, the Gran Chaco has an abundance of wildlife. Among the larger animals are the jaguar, ocelot, puma, tapir, giant armadillo, spiny anteater, many foxes, numerous small wildcats, the agouti (a large rodent), the capybara (water hog), the maned wolf, the palustrian deer, the peccary, and the guanaco (a camelid related to the llama). The Chaco is one of the last major refuges for the rhea (or nandu), a large, flightless South American bird, and it has long been noted for its abundant and varied bird population. The streams are host to more than 400 fish species, among which are the salmonlike dorado and the flesh-eating piranha. Countless travelers' tales complain of the pestilent insects. Reptiles also are abundant, with numerous lizards and at least 60 known species of snakes, including many pit vipers and constrictors, while at least six species of poisonous tree toads have been identified.

The people and economy. *Early settlement.* The indigenous peoples of the Chaco were numerous. Because of their subsistence as hunters, gatherers, and fishermen, tribal units were not much larger than extended families. Nevertheless, from among the diverse dialects, anthropologists have described a few major linguistic associations: the Guaycurú, Lengua, Mataco, Zamuco, and Tupí-Guaraní. Most of these people lived under extremely primitive conditions; settlement depended on the availability of fresh water, making stream courses the most coveted sites. Implements were fashioned largely from wood and bones because of the absence of stones, while the spiny leaves of the pineapple-like groundcover *carraguatas* served as a universal source of fibre. The Chaco forest, despite its harshness, contained more plant sources of human sustenance—*e.g.,* edible pods, fruits, berries, and tubers—than surrounding areas, and this factor was well exploited by the native peoples. Game was gathered by trapping, netting, clubbing, and spearing, often in conjunction with large group drives. For those Indian groups still living outside the limits of European settlement, conditions are only slightly modified today, although these people now have domesticated animals and metal tools. Most tribes, however, exist as sort of a peasant pioneer fringe and practice some form of shifting subsistence agriculture.

European colonization and economic activity. Aside from the scattered (although successful) agricultural communes (*reducciones*) of the Jesuits and the settlement of Asunción, Paraguay, on its eastern fringe, the Chaco defied effective European occupancy until well into the 19th century. Hostile Indian groups, in concert with the forbidding nature of the Chaco itself, limited European influence in the colonial period to a situation much like a state of siege.

The limited early colonization in the Argentine and Bolivian Chaco was based on exploitation of the longhorn criollo (or Creole) cattle that roamed half wild throughout the region. The western Chaco Austral, near Salta, also was exploited as a source of heavy timbers for the mines in the highlands of Bolivia and Peru. In the late 19th century, the Chaco in Argentina and southern Paraguay became a land of great ranches (*estancias*) raising criollo cattle, and numerous, small, independent camps (*obrajes*) of woodcutters exploited the abundant hardwoods of the Chaco forests for lumber and firewood. Cattle grazing has continued to be the most extensive use of the land, with few substantial changes from pioneer days. One of the key problems in improving the cattle industry has been the apparent endemic nature of many serious cattle diseases and pests against which criollo cattle have developed some immunity, whereas purebred cattle have remained fully susceptible.

Cattle raising and logging

In the eastern Chaco, vast, highly capitalized industrial ventures established large plants to process the great quantities of tannin found in the various quebracho species. Unlike the *obrajes* of the woodcutters, these operations were large, centralized mills adjacent to rivers or rail lines, from which the selective cutting of quebracho has proceeded at a systematic pace. The slow growth habits of the quebracho trees, however, pose a threat to the tannin industry, as the pace of the harvest easily can exceed reforestation efforts. The relatively untouched Bolivian Chaco contains stands of quebracho timber, but most of these are in remote areas and have not been exploited while production has continued in the more accessible areas. Quebracho tannin has remained one of the economic mainstays of the Chaco, but it has faced competition from other sources of tannin, both natural and synthetic. Other forest products include lumber and heavy timbers from a variety of other species, firewood, and palo santo oil from the wood of *Bulnesia sarmientii,* a tree found in the more arid portions of the Chaco.

Modern developments. Cotton has become another principal crop of the Chaco. Wild cotton has been known in much of the region since pre-Columbian times, but it never was grown as anything more than an agricultural curiosity until the 20th century. During World War I, with cotton prices at a peak, large areas in Argentina's Chaco province were converted to cotton cultivation. Production was enhanced considerably by the use of irrigation and the development of drought-resistant stocks. The crop area subsequently has been expanded in Argentina, and cotton has become important in Paraguay; it also is grown in lesser quantities in Bolivia. These increases have occurred despite bad markets, insect plagues, and often poor weather in many years and more recent problems with soil erosion. Both fibre and cottonseed oil are produced, mainly for domestic consumption. Other crops include linseed, sunflowers, sorghum, and corn (maize).

The discovery of oil in the Bolivian piedmont in the 1920s led within a decade to the disastrous Chaco War between Bolivia and Paraguay, each country hoping to find more oil in the neighbouring Chaco Boreal. Paraguayan claims eventually were honoured, but they did not include any part of the oil-rich piedmont; subsequent explorations in Paraguay have been disappointing. Oil has been discovered across the border in Argentina, however, and large quantities of natural gas have been recovered on the northern fringe of the Bolivian Chaco near Santa Cruz.

Since World War II, efforts have been made by the respective governments to spur settlement of the Chaco. Argentine interest has been concentrated along the railways out of Resistencia and Formosa, with pioneer settlements composed mainly of eastern European immigrants and based on cotton production. In the central Paraguayan Chaco, which has been accessible by road only since 1965, Mennonite immigrants from Canada had settled in the 1920s and were joined by coreligionists from the Soviet Union in the 1930s. These settlers established self-sufficient colonies and were joined by another large contingent of refugees from the Soviet Union after World War II.

The Mennonite colonies

The Southern and Central Andes and Patagonia.

The primary land use in the Bolivian Chaco is still open cattle range. The nearby supplies of oil and natural gas and the hydroelectric and water storage capacity of such fast-flowing piedmont streams as the Pilcomayo, however, offer great development potential.

(G.E.Ma./M.D.H.M./K.E.W.)

PATAGONIA

The region of Patagonia covers nearly all of the southern portion of mainland Argentina. With an area of about 260,000 square miles (673,000 square kilometres), it constitutes a vast area of steppe and desert that extends from latitude 37° to 51° S. It is bounded, approximately, by the Patagonian Andes to the west, the Colorado River to the north (except where the region extends north of the river into the Andean borderlands), the Atlantic Ocean to the east, and the Strait of Magellan to the south; the region south of the strait—Tierra del Fuego, which is divided between Argentina and Chile—also is often included in Patagonia.

The name Patagonia is said to be derived from Patagones, as the Tehuelche Indians, the region's original inhabitants, were called by 16th-century Spanish explorers. According to one account, Ferdinand Magellan, the Portuguese navigator who led the first European expedition into the area, coined that name because the appearance of the Tehuelche reminded him of Patagon, a dog-headed monster in the 16th-century Spanish romance *Amadís of Gaul.*

Physical features. *Physiography.* Desert and semidesert cover the Patagonian tableland that extends from the Andes to the Atlantic Ocean. The general aspect of this tableland is one of vast steppelike (*i.e.,* virtually treeless) plains, rising in terrace fashion from high coastal cliffs to the foot of the Andes; but the true aspect of the plains is by no means as simple as such a general description would imply. The land along the Negro River rises in a series of fairly level terraces from about 300 feet (90 metres) at the coast to about 1,300 feet at the junction of the Limay and Neuquén rivers and 3,000 feet at the base of the Andes. The tableland region rises to an altitude of 5,000 feet.

South of the Negro River, the plains are much more irregular. Volcanic eruptions occurred in this area until fairly recent times, and basaltic sheets covered the tableland east of Lakes Buenos Aires and Pueyrredón. Near the Chico and Santa Cruz rivers, the plains have spread to within about 50 miles (80 kilometres) of the coast and reach almost to the coast south of the Coig and Gallegos rivers. In places, basaltic massifs (mountain masses) are the salient features of the landscape.

The coast

The coast consists largely of high cliffs separated from the sea by a narrow coastal plain. Thus, the plateaus are formed of horizontal strata, some of sedimentary rocks and others of lava flows. Areas of hilly land, composed of resistant crystalline rocks, stand above the plateaus.

Drainage and soils. The deep, wide valleys bordered by high cliffs that cut the tablelands from west to east are all beds of former rivers that flowed from the Andes to the Atlantic; only a few now carry permanent streams of Andean origin (the Colorado, Negro, Chubut, Senguerr, Chico, and Santa Cruz rivers). Most of the valleys either have intermittent streams—such as the Shehuen, Coig, and Gallegos rivers, which have their sources east of the Andes—or contain streams like the Deseado River, which completely dry up along all or part of their courses and are so altered by the combined effect of wind and sand as to afford little surface evidence of the rivers that once flowed in them. Still other streams, such as the Perdido, terminate in basins containing salt flats or salt ponds. The canyon bottoms consist mostly of deep beds of coarse alluvial sands and gravels, which act as groundwater reservoirs to supplement the scanty surface water.

The line of contact between the Patagonian tableland and the Patagonian Andes is marked by a chain of lakes found in glacier troughs or cirques that are dammed downslope by moraines and other glacial landforms consisting of unconsolidated and unsorted till. From Lake Nahuel Huapí northward, the lakes—except for Lake Lácar—drain to the Atlantic. South of Lake Nahuel Huapí, however, all the lakes except Viedma and Argentino drain to the Pacific through deep canyons that have been cut from west to east across the cordillera by headward erosion.

The best soils in Patagonia are found north of the Negro River, especially where they are formed from volcanic rock. Proceeding south, the soils become increasingly arid and stony, and broad expanses of stream-rounded pebbles, called *grava patagónica,* often are found on level ground.

Climate. Patagonia is influenced by the South Pacific westerly air current, which brings humid winds from the ocean to the continent. These winds, however, lose their humidity (through cooling and condensation) as they blow over the west coast of South America and over the Andes, and they are dry when they reach Patagonia. Patagonia can be divided into two main climatic zones—northern and southern—by a line drawn from the Andes at about latitude 39° S to a point just south of the Valdés Peninsula, at about 43° S.

The two climatic zones

The northern zone is semiarid, with annual mean temperatures between about 54° and 68° F (12° and 20° C); recorded maximum temperatures vary from about 106° to 113° F (41° to 45° C), and minimum temperatures from 12° to 23° F (−11° to −5° C). Sunshine, minimal along the coast, is most plentiful inland to the northwest. Annual rainfall amounts vary from about 3.5 to 17 inches (90 to 430 millimetres). The prevailing winds, from the southwest, are dry, cold, and strong.

The climate of the southern zone is sharply distinct from the humid conditions of the Andean cordillera to the west. In the northern part of the zone, Atlantic influences are practically nonexistent—probably because of the relatively high elevations of the coastal region, which reach 900 to 1,800 feet around San Jorge Gulf—although cold Pacific winds from the west and the cold Falkland Current off the Atlantic coast do have some effect. In the southern part, which becomes increasingly peninsular with higher latitude, the Atlantic exerts some influence. The zone has a cold, dry climate, with temperatures that are higher along the coast than they are inland and with strong west winds. Mean annual temperatures range from 40° to 55° F (4° to 13° C), with the maximum temperature reaching about 93° F (34° C) and minimum temperatures between 16° and −27° F (−9° and −33° C). Heavy snows fall in winter, and frosts can occur throughout the year; spring and autumn provide only short transitions between summer and winter. Average annual precipitation (rain and snow) ranges between about 5 and 8 inches, though as much as 19 inches has been recorded. Less precipitation falls in the arid central areas, which also receive more sunshine than the coast or the Andean cordillera.

Plant life. The long, narrow strip of Patagonia's western border supports vegetation like that found in the adjacent cordillera, primarily deciduous and coniferous forests. The vast tableland region is divided into northern and southern zones, each of which has its own characteristic vegetation.

The larger northern steppe zone extends south to about latitude 46° S. In the north is found *monte* vegetation—xerophytic (drought-tolerant) scrub forests—which gives way farther south to open bushland of widely spaced thickets between about 3 and 7 feet high. Grasses flourish in the sandy areas, while halophytic (salt-tolerant) grasses and shrubs predominate in the salt flats. The southern, more arid, zone extends south of 46° S. The vegetation is low and considerably more sparse and needs almost no water.

Animal life. Among the Patagonian birds are herons and other waders; predators such as the shielded eagle, the sparrow hawk, and the chimango (or beetle eater); and the almost extinct rhea (nandu). The typical marsupial of the region is the *comadreja* (a member of the weasel family). Species of bats include a long-eared variety. Armadillos, pichis (small armadillos), foxes, ferrets, skunks, mountain cats, and pumas are to be found, as are the Patagonian cavy (or mara) and different kinds of burrowing rodents, such as the vizcacha and the tuco-tuco. Of the larger mammals, the most noteworthy is the guanaco, a kind of llama, which has been hunted almost to extinction.

Patagonia has a number of species of poisonous snakes, as well as tortoises and a variety of lizards. Among the arthropods and arachnids are vinchucas (winged bugs), bloodsucker insects (transmitters of American trypanoso-

Herd of guanacos in eastern Patagonia, Valdés Peninsula, Argentina.

miasis, or Chagas' disease), scorpions, and several kinds of spiders, including one endemic genus called *Mecysmanchenius*. The rivers and lakes are naturally poor in fish, but some have been stocked with salmon and trout. Marine fish, however, as well as crustaceans and mollusks are plentiful off the coast.

Oil and natural gas

The economy. *Resource exploitation.* The oil fields around Comodoro Rivadavia and near Neuquén contain most of Argentina's reserves, and natural gas also has been found in these two areas; these are Patagonia's most valuable mineral assets. In addition, deposits of iron ore are worked at Sierra Grande, and some coal is mined in the south near Río Turbino. Other mineral deposits include manganese, tungsten (wolframite), fluorite (calcium fluoride), lead, heavy spar (barite, the principal ore of barium), copper and gold, vanadium, zinc-lead ore, and uranium. There also are deposits of kaolin and gypsum.

Dams have been constructed on the Neuquén and Limay rivers in order to exploit the hydroelectric potential of the western portion of Patagonia. The projects also have created large reservoirs that have made extensive irrigated agriculture possible in the Negro River region. Among the major crops grown are peaches, plums, almonds, apples, pears, olives, grapes, hops, dates, vegetables, aromatic plants, and alfalfa.

Tourism has become important since the end of World War II, as wildlife reserves and the national parks located along the Patagonian Andes have brought in growing numbers of those seeking recreation. There also has been an increase in scientific study (*e.g.*, glacier research) and in detailed mapping and surveying for mineral exploitation.

Transportation. Comodoro Rivadavia is connected to Buenos Aires by a road that runs more than 1,860 miles through the Patagonian coastal region. The roads farther inland, however, are few and of poor quality. Several railroads traverse the region from east to west; two that reach the foothills of the Andes are connected to Buenos Aires. Air services are focused chiefly on the towns of the coastal region. The chief ports are Rawson, Deseado, and Río Gallegos; San Antonio Oeste and Puerto Madryn, both on protected bays and developed for international traffic; and Comodoro Rivadavia, an outlet for petroleum products.

History. The original inhabitants of Patagonia consisted mostly of Tehuelche Indians, who are thought to have come from Tierra del Fuego. The most ancient artifacts, such as harpoons, found in the caves along the Strait of Magellan suggest that these people were moving up the mainland coast about 5,100 years ago. The robust and tall Tehuelche were divided into northern and southern groups, each with its own dialect. Spanish explorers found the Tehuelche living as nomadic hunters of guanaco and rhea. The surviving descendants of these people are few in number, nearly all of them having been assimilated into Spanish culture.

Early European settlement

Toward the end of the 16th century, the Spaniards attempted to colonize the Patagonian coastal region to clear it of English pirates, but a Jesuit settlement on San Matías Gulf came to nothing. In 1778 the English tried to settle on the same bay, and the Spaniards reacted by founding Patagonia's first two towns, San José and Viedma (originally named Nuestra Señora del Carmen). A Spanish settlement at Puerto Deseado lasted from 1780 to 1807, but three years later this region again was devoid of European settlement.

After Argentina became independent, Patagonia largely was left alone, until it was cleared of Indian occupation in the Conquest of the Desert campaigns of the 1870s. An attempt was then made to settle the region and to make it part of the national state. Immigration, however, was not massive, though people came for various reasons: some to exploit the economic resources and others (*e.g.*, the Welsh) to enjoy religious or political liberties. The mineral wealth of the region in particular attracted immigrants from Chile, and Chileans seeking temporary work rather than a fixed domicile now constitute the largest proportion of the population. Apart from major concentrations at Comodoro Rivadavia and in the towns strung out along the upper valley of the Negro River, Patagonia's sparse population is mostly rural. (E.F.G.D./K.E.W.)

Drainage systems

AMAZON RIVER BASIN

The Amazon (Portuguese and Spanish: Amazonas) is the greatest river of South America and the largest drainage system in the world in terms of the volume of the river's flow and the area of its basin. The total length of this great river—measured from the headwaters of the Ucayali-Apurímac river system in Peru—is about 4,000 miles (6,400 kilometres), which is slightly shorter than the Nile River but still the equivalent of the distance from New York City to Rome. Its westernmost source is high in the Andes Mountains, within 100 miles of the Pacific Ocean, and its mouth is in the Atlantic Ocean.

Extent of the basin

The vast Amazon basin (Amazonia), the largest lowland in Latin America, has an area of about 2.3 million square miles (6 million square kilometres) and is nearly twice as large as that of the Congo River, the Earth's other great equatorial drainage system. Stretching some 1,725 miles from north to south at its widest point, the basin includes the greater part of Brazil and Peru, significant parts of Colombia, Ecuador, and Bolivia, and a small area of Venezuela; roughly two-thirds of the Amazon's main stream and by far the largest portion of its basin are within Brazil. The Tocantins-Araguaia catchment area in Pará state covers another 300,000 square miles. Although considered a part of Amazonia by the Brazilian government and in popular usage, it is technically a separate system. It is estimated that about one-fifth of all the water that runs off the Earth's surface is carried by the Amazon. The flood-stage discharge at the river's mouth is about 6,180,000 cubic feet (175,000 cubic metres) per second, which is four times that of the Congo and more than 10 times the amount carried by the Mississippi River. This immense volume of fresh water dilutes the ocean's saltiness for more than 100 miles from shore.

The Central and Northern Andes and the Amazon River basin and its drainage network.

The extensive lowland areas bordering the main river and its tributaries, called *várzeas*, are subject to annual flooding, with consequent soil enrichment; however, most of the vast basin consists of upland, well above the inundations and known as *terra firme*. More than two-thirds of the basin is covered by an immense rain forest, which grades into dry forest and savanna on the higher northern and southern margins and into montane forest in the Andes to the west. The Amazon Rain Forest, which represents about half of the Earth's remaining rain forest, also constitutes its largest reserve of biological resources.

The first European to explore the Amazon, in 1541, was the Spanish soldier Francisco de Orellana, who is said to have given the river its name after reporting pitched battles with tribes of female warriors, whom he likened to the Amazons of Greek mythology. Although the name Amazon is conventionally employed for the entire river, in Peruvian and Brazilian nomenclature it properly is applied only to sections of it. In Peru the upper main stream (fed by numerous tributaries flowing from sources in the Andes) down to Iquitos (Peru) is called Marañón (Portuguese: Maranhão), and from there to the Atlantic it is called Amazonas. In Brazil the name Solimões is used from Iquitos to the mouth of the Negro River and Amazonas only from the Negro to the sea.

Physical features. *Landforms and drainage patterns.* The Amazon basin is a great structural depression, a subsidence trough that has been filling with immense quantities of sediment of Cenozoic age (*i.e.*, from the past 66.4 million years). This depression, which flares out to its greatest dimension in the Amazon's upper reaches, lies between two old and relatively low crystalline plateaus, the rugged Guiana Highlands to the north and the lower Brazilian Highlands (lying somewhat farther from the main river) to the south. The Amazon basin was occupied by a great freshwater sea during the Pliocene Epoch (5.3 to 1.6 million years ago). Sometime during the Pleistocene Epoch (1,600,000 to 10,000 years ago) an outlet to the Atlantic was established, and the great river and its tributaries became deeply entrenched into the former Pliocene seafloor.

The modern Amazon and its tributaries occupy a vast system of drowned valleys that have been filled with alluvium. With the rise in sea level that followed the melting of the Pleistocene glaciers, the steep-sided canyons that had been eroded into the Pliocene surface during the period of lower sea levels were gradually flooded. In the upper part of the valley—in eastern Colombia, Ecuador, Peru, and Bolivia—more recent outwash from the Andes has covered many of the older surfaces.

Physiography of the river course. The Amazon River has its main outlet north of Marajó Island, a lowland somewhat larger in size than Denmark, through a cluster of half-submerged islets and shallow sandbanks. Here the mouth of the river is 40 miles (64 kilometres) wide. The port city of Belém is on the deep water of the Pará River, an estuary marking the south side of Marajó, which is fed chiefly by the Tocantins River entering it southwest of Belém. The port city's link with the main Amazon channel is either north along the ocean frontage of Marajó or following the deep but narrow *furos* (channels) of Breves that bound the island on the southwest and link the Pará River with the Amazon. There are more than 1,000 tributaries of the Amazon that flow into it from the Guiana Highlands and from the Brazilian Highlands, as well as from the Andes. Seven of these tributaries—the Japurá (Caquetá in Colombia), Juruá, Madeira, Negro, Purus, Tocantins, and Xingu rivers—are more than 1,000 miles long; and one, the Madeira River, exceeds 2,000 miles from source to mouth. The largest oceangoing ships can ascend the river 1,000 miles to the city of Manaus, while lesser freight and passenger vessels reach Iquitos, Peru, 1,300 miles farther upstream, at any time of year.

Major tributaries

The sedimentary axis of the Amazon basin comprises two distinct groups of landforms: the *várzea*, or floodplain of alluvium of Holocene age (*i.e.*, up to 10,000 years old), and the *terra firme*, or upland surfaces of Pliocene and Pleistocene materials (those from 10,000 to 5,300,000 years old) that lie well above the highest flood level. The floodplain of the main river is characteristically 12 to 30 miles wide. It is bounded irregularly by low bluffs 20 to 60 feet high, beyond which the older, undulating upland extends both north and south to the horizon. Occasionally these bluffs are undercut by the river as it swings to and fro across the alluvium, producing the *terra caída*, or "fallen land," so often described by Amazon travelers. At the city of Óbidos, where the river narrows to a width of 1.25 miles, a low range of relatively hard rock interrupts the otherwise continuous floodplain.

The streams that rise in the ancient crystalline highlands—the Jari, Trombetas, and Negro to the north and the Tocantins-Araguaia, Xingu, and Tapajós to the south—are so-called "blackwater" streams; they are acidic and rich in humus. Because these streams originate in nutrient-poor, often sandy uplands, they carry little or no silt or dissolved solids. Where such blackwater tributaries enter the main river, they are sometimes blocked off to form funnel-shaped, freshwater lakes or estuaries, as at the mouth of the Tapajós.

Blackwater rivers

In contrast, the Madeira River, which joins the Amazon some 50 miles downstream from Manaus and its principal affluents—the Purus, Juruá, Ucayali, and Huallaga on the right or southern bank and the Japurá (Caquetá), Içá (Putumayo), and Napo from the northwest—have their source in the youthful and tectonically active Andes. There they pick up the heavy sediment loads that account for their "whitewater" designation. Where the silt-laden waters of the Amazon (Solimões in Brazil), derived from these streams, meet those of the Negro at Manaus, the darker and hence warmer and sediment-free waters of the latter tend to be overrun by those of the Amazon, creating a striking colour boundary which is erased by turbulence downstream.

The mother river, the Marañón above Iquitos, rises in the central Peruvian Andes at an elevation of 15,870 feet in a small lake in the Cordillera Huayhuash above Cerro de Pasco. The Huallaga and Ucayali, major right-bank affluents of the Marañón, originate considerably farther south. The headwaters of the deeply entrenched Apurímac and Urubamba, tributaries in turn of the Ucayali, reach to within 100 miles of Lake Titicaca (elevation 12,500 feet) on the Peru-Bolivia border, the farthest of any stream in the system from the great river's mouth.

The Negro River, the largest of all the Amazon tributaries, accounts for about one-fifth of the total discharge of the Amazon and 40 percent of its aggregate volume measured just below the confluence at Manaus. Its drainage area of 292,000 square miles includes that of the Branco, its major left-bank tributary, with its source in the Guiana Highlands. Another of the Negro's affluents, the Casiquiare, is a product of the bifurcation of the Orinoco River; it forms a link between the Amazon and the Orinoco's drainage system. The Branco watershed, approximately coincident with the state of Roraima, includes extensive tracts of sandy, leached soils that support a grassy and stunted arboreal cover (*campos*). Other tributaries of the Negro, such as the Vaupés and Guainía, drain eastward from the Colombian Oriente. The river traverses some of the least populous and least disturbed parts of the Amazon basin, including several national parks, national forests, or indigenous reserves. In its lower reaches it becomes broad and island-filled, in places reaching widths of 20 miles.

The Negro River and its affluents

The Madeira River, second largest affluent of the Amazon, has a discharge of perhaps two-thirds that of the Negro. Silt from its turbid waters has choked its lower valley with sediments; where it joins the Amazon below Manaus, it has contributed to the formation of the 200-mile-long island of Tupinambarana. Beyond its first cataract 600 miles up the river, its three major affluents—the Madre de Dios, the Beni, and the Mamoré—gave easy access to the rubber-rich forests of the Bolivian Oriente, while the Mamoré's tributary, the Guaporé, opened the way to the goldfields of Mato Grosso. Even more important to the rubber tappers were the meandering Purus and Juruá rivers that flank the Madeira on the west.

Hydrology. Most of the estimated 1.3 million tons of sediment that the Amazon pours daily into the sea is transported northward by coastal currents to be deposited along the coasts of northern Brazil and Guiana. As a

consequence, the river is not building a delta. Normally, the effect of the tide is felt as far upstream as Óbidos, 600 miles from the river's mouth. A tidal bore called the *pororoca* occurs at times in the estuary prior to spring tides. With an increasing roar it advances upstream at 10 to 15 miles per hour, forming a breaking wall of water from 5 to 12 feet high.

At the Óbidos narrows, the flow of the river has been measured at 216,000 cubic metres per second; its width is constricted to little more than a mile. Here the average depth of the channel below the mean watermark is more than 200 feet, well below sea level; in most of the Brazilian part of the river its depth exceeds 150 feet. Its gradient is extraordinarily slight. At the Peruvian border, some 2,000 miles from the Atlantic, the elevation above sea level is less than 300 feet. The maximum free width (without islands) of the river's permanent bed is 8.5 miles, upstream from the mouth of the Xingu. During great floods, however, when the river completely fills the floodplain, it spreads out in a band 35 miles wide or more. The average velocity of the Amazon is about 1.5 miles per hour, a speed that increases considerably at flood time.

The Amazon floods

The rise and fall of the water is controlled by events external to the floodplain. The floods of the Amazon are not disasters but rather distinctive, anticipated events that define the calendar and the rhythm of life much as seasons do elsewhere. Their marked regularity and the gradualness of the change in water level are due to the enormous size of the basin, the gentle gradient, and the great temporary storage capacity of both the floodplain and the estuaries of the river's tributaries. The upper course of the Amazon has two annual floods, and the river is subject to the alternate influence of the tributaries that descend from the Peruvian Andes (where rains fall from October to January) and from the Ecuadorian Andes (where rains fall from March to July). This pattern of alternation disappears farther downstream, the two seasons of high flow gradually merging into a single one. Thus, the rise of the river progresses slowly downstream in a gigantic wave from November to June, and then the waters recede until the end of October. The flood levels are, in places, from 40 to 50 feet above low river.

Climate. The climate of Amazonia is warm, rainy, and humid. The length of day and night is equal on the Equator (which runs only slightly north of the river), and the usually clear nights favour relatively rapid radiation of the heat received from the sun during the 12-hour day. There is a greater difference between daytime and midnight temperatures than between the warmest and coolest months. Hence, night is the winter of the Amazon. At Manaus, the average daily temperature is 89° F (32° C) in September and 75° F (24° C) in April, but the humidity is consistently high and often oppressive. During the winter months of the Southern Hemisphere, a powerful south-polar air mass occasionally pours northward into the Amazon region, causing a sharp drop in temperature, known locally as a *friagem,* when the mercury may register as low as 57° F (14° C). At any time of the year, several days of heavy rain can be succeeded by clear, sunny days and fresh, cool nights with relatively low humidity. In the lower reaches of the river basin, cooling trade winds blow most of the year.

The main influx of atmospheric water vapour into the basin comes from the east. About half of the precipitation that falls originates from the Atlantic Ocean; the other half comes from evapotranspiration from the tropical forest and associated convectional storms. Rainfall in the lowlands typically ranges from 60 to 120 inches (1,500 to 3,000 millimetres) annually in the central Amazon basin (*e.g.,* Manaus). On the eastern and western margins of the

Rainfall

basin, rainfall occurs throughout the year, whereas in the central part there is a definite drier period, usually from June to November. Manaus has experienced as many as 60 consecutive days without rain, but such droughty periods are uncommon. The dry season is not sufficiently intense to arrest plant growth, but it may facilitate the onset and spread of fires, whether arsonous or natural. To the west the Andes form a natural barrier that prevents most of the water vapour from leaving the basin.

Along the southern margin of the Amazon basin the climate grades into that of west-central Brazil, with a distinct dry season during the Southern Hemispheric winter. As elevations increase in the Andes, temperatures fall significantly.

Soils. The vast Amazonian forest vegetation appears extremely lush, leading to the erroneous conclusion that the underlying soil must be extremely fertile. In fact, the nutrients in the system are locked up in the vegetation, including roots and surface litter, and are continuously recycled through leaf fall and decay. Generally, the soils above flood level are well-drained, porous, and of variable structure. Often they are sandy and of low natural fertility because of their lack of phosphate, nitrogen, and potash and their high acidity. Small areas are underlain with basaltic and diabasic rocks, with reddish soils (*terra roxa*) of considerable natural fertility. The *terra preta dos Indios* ("black earth of the Indians") is another localized and superior soil type.

The agricultural potential of the annually flooded *várzea* areas is great. Their soils do not lack nutrients, since they are rejuvenated each year by the deposit of fertile silt left as the waters recede, but use for agricultural purposes is limited by the periodic inundations. It is estimated that these valuable soils occupy some 25,000 square miles.

The rain forest

Plant life. The overwhelmingly dominant feature of the Amazon basin is the tropical rain forest, or selva. From the air the Amazon forest appears to stretch unbroken to the horizon like a tufted green carpet. Closer inspection reveals its bewildering complexity and prodigious variety of trees; as many as 100 arboreal species have been counted on a single acre of forest with hardly any one of them occurring more than once.

The Amazon forest has a strikingly layered structure. The sun-loving giants of the uppermost story, the canopy, soar to as much as 120 feet above the ground; occasional individual trees, known as emergents, rise beyond the canopy, frequently attaining heights of 200 feet. Their straight, whitish trunks are splotched with lichens and fungus. A characteristic of these giant trees is the buttresses, or basal enlargements of their trunks, which presumably help stabilize the top-heavy trees during infrequent heavy winds. Further characteristics of the canopy trees are their narrow, downward-pointing "drip-tip" leaves that easily shed water and their cauliflory (the production of flowers directly from the trunks rather than from the branches). Flowers are inconspicuous. Among the canopy species, which capture most of the sunlight and conduct most of the photosynthesis, prominent members include the rubber tree (*Hevea brasiliensis*), the silk-cotton (*Ceiba pentandra*), the Brazil nut (*Bertholletia excelsa*), the sapucaia (*Lecythis*), and the sucupira (*Bowdichia*). Many creatures, including monkeys and sloths, spend their entire lives in this sunlit canopy. Below it are found two or three levels of shade-tolerant trees, including many species of palms, such as *Mauritia, Orbignya,* and *Euterpe.* Myrtles, laurels, bignonias, figs, Spanish cedars, mahogany, and rosewoods are also common. They support a myriad of epiphytes (plants living on other plants)—such as orchids, bromeliads, and cacti—as well as ferns and mosses. The entire system is laced together by a bewildering network of woody ropelike vines known as lianas.

In addition to the rain forests of the *terra firme,* there are two types of inundated rain forests, *várzea* and *igapó,* which constitute about 3 percent of the total Amazonian rain forest. *Várzea* forests can be found in the silt- and nutrient-rich floodplains of whitewater rivers such as the Madeira and the Amazon, with their ever-changing mosaic of lakes, marshes, sandbars, abandoned channels, and natural levees. They are generally not as high, diverse, or old as those of the *terra firme,* being subject to periodic destruction by floods and human manipulation. (The *várzea* and its flood-free margins are the principal rainforest habitat of human beings.) Wild cane (*Gynerium*) and aquatic herbs and grasses, as well as fast-growing pioneer tree species such as *Cecropia, Ficus,* and *Erythrina,* are conspicuous.

Igapó forests grow along the sandy floodplains of blackwater rivers such as the Negro, the Tapajós, and the Trom-

betas. Because human settlement is limited in these plains, there may be undisturbed, seasonally flooded forests that stand in water for up to half the year, the water reaching heights of up to 40 feet. A canoe can often be paddled between the trunks of trees adapted to such an aquatic environment.

The lowland rain forest on the Andean fringe grades into a discontinuous, tangled montane or cloud forest of misshapen trees cloaked with mosses, lichens, and bromeliads. Here one encounters the cinchona or fever-bark tree, once exploited for its antimalarial agent quinine. At still higher elevations is found the grass and shrub growth of the cold *puna* and *páramo* regions.

Along the drier, southern margin of the Amazon basin, high forest gives way to the *cerrado* (savanna and scrub) and *caatinga* (heath forest). The latter is characteristic of parts of the Mato Grosso Plateau, where taller forest is restricted to the stream courses and swales (marshy depressions) that dissect the upland surface. On the sandy soils of the lower Negro and the Branco drainage areas and locally in Amapá, grassy savannas dotted with stunted trees replace the high forest.

Animal life. To give a succinct overview of the complete fauna of the Amazon is as impossible as it is to adequately describe the great diversity of its flora; in part this is because many of the region's species have yet to be identified. The rivers and streams of the basin teem with life, and the forest canopy resonates with the cries of birds and monkeys and the whine of insects. There is a notable paucity of large terrestrial mammal species; indeed, many of the mammals are arboreal.

More than 8,000 species of insects alone have been collected and classified. Myriads of mosquitoes plague travelers and may transmit such diseases as malaria and yellow fever. Leaf-cutting ants (*Atta* and *Acromyrmex*) and other pests may torment the traveler. The most troublesome insects of all are the ubiquitous, small, black flies, called *piums*, whose bite can itch for days.

The Amazon and its tributaries, together with the bordering *várzea* lakes and flooded forests, constitute a vast sea of fresh water, much of it slowly flowing, which teems with fish life. About 1,500 fish species have been found within the Amazon system, but many more remain unidentified. Most fish are migratory, moving in great schools at spawning time. Fish represent a critical source of protein in the often meat-poor diet of the peasant (*caboclo*) population (the term *caboclo* is used for the peasant population of mixed Indian-European blood). Among the more important commercial species are the pirarucu (*Arapaima gigas*), one of the world's largest freshwater fish, and various giant catfish. The well-known, small, flesh-eating piranha generally feeds on other fish but may attack any animal, including humans, that enters the water; its razor-sharp teeth cut out chunks of flesh, stripping a carcass of its meat in a few minutes. The traffic in frozen and dried fish to urban markets has increased to such a degree that some stocks are locally threatened. With the rapid means of transport afforded by jet airplanes, a worldwide market has developed for tropical aquarium fish distinctive to the Amazon. Iquitos, Manaus, and the Colombian port of Leticia are centres of this trade.

Fish

Crocodiles are hunted for their skins; river turtles and their eggs are considered a delicacy; the giant sea cow, or manatee, is sought for its flesh and for oil. All are threatened by overhunting, and the manatee has been listed as an endangered species. Aquatic animals also include freshwater dolphins (*Inia geoffrensis*); the capybara, the largest rodent in the world (weighing up to 170 pounds); and the nutria, or coypu, valued especially for its pelt. Other common rodents are the paca, agouti, porcupine, and local species of squirrels, rats, and mice.

The tapir, the white-lipped peccary, and several species of deer are native to the Amazon basin and much sought for their meat. Water buffalo, introduced from Southeast Asia as work and dairy cattle, have run wild in the remote, swampy parts of Marajó Island.

Especially characteristic of the Amazon forest are several species of monkeys. Of note are the howler monkeys, which make the selva resound with their morning and evening choruses. The small, agile squirrel monkey, the most ubiquitous of Amazonia's monkeys, is used in laboratories, as is the larger spider monkey. Among a host of other primate species are woolly monkeys, capuchin monkeys, titis, sakis, and marmosets. All species are used for food and frequently are seen for sale in local markets. As the human population increases and the shotgun replaces the blowgun, pressure on the wild fauna is mounting.

Large cats, such as the jaguar and ocelot, are rare, although pumas may be found in larger numbers in the Andean fringe of the basin. Smaller carnivores include coati, grisons, and weasels. Countless bats inhabit the Amazonian night, including the blood-drinking vampire bat.

Large cats

The Amazon basin is exceedingly rich in birdlife. Morning and evening, the parrots and macaws fly to and from their feeding grounds, their brilliant plumage flashing in the sunlight and their raucous voices calling out their presence. Through the day the caciques quarrel in trees where their hanging nests swing by the dozens. Hoatzins screech in noisy flocks from streamside brush, while solitary hawks and eagles scream from tree stumps. Everywhere is heard the twittering of small birds, the sound of woodpeckers, and the gutteral noises of such waterbirds as herons, cormorants, roseate spoonbills, and scarlet ibises. Parakeets, more common than sparrows in the United States, fly around in great flocks. At dusk, toucans cry a discordant plaint from the treetops and are joined by ground-dwelling tinamous and quail. The night air is filled with the cries of various species of nightjar.

The people. *Early settlement patterns.* At the time of the European conquest, the bottomlands and fringing upland surfaces of the Amazon River and its major tributaries supported relatively dense, sedentary populations of indigenous peoples who practiced intensive root-crop farming, supplemented by fishing and by hunting aquatic mammals and reptiles. The higher areas away from the rivers and their floodplains, were—and still are in some of the more remote sectors—inhabited by small, widely dispersed, seminomadic tribes of Indians. These groups traditionally have relied predominantly on hunting large and small animals and on gathering wild fruits, berries, and nuts, while practicing some small-patch agriculture of low yield. In the early 1990s the Indian population of the Amazon basin numbered about 600,000, of whom perhaps close to one-third live in Brazil and the rest in the Oriente of the four Andean countries.

Indians

The Amazonian Indians early devised means of making the poisonous bitter cassava (manioc) edible; the end product, called *farinha*, became a food staple widely used today in much of tropical America. Amazonian Indians perfected the use of quinine as a specific against malaria, extracted cocaine from the leaves of the coca tree, and collected the sap of the Brazilian rubber tree (*Hevea brasiliensis*). They were skilled navigators in their dugout canoes and sailing rafts (*jagandas*), and they invented the blowgun and the hammock. One of their ancient arrow poisons, curare (*Chondrodendron tomentosum*), has been used in modern times in the therapy of a host of paralyses and spastic disorders, such as multiple sclerosis.

The early European explorers of the Amazon provisioned themselves from the food supplies of the Indians they met and commandeered their canoes. Large numbers of Indians were taken into slavery, especially during the organized raids (*bandeiras*) of the 16th to 18th century; many others succumbed to such European diseases as influenza, measles, and smallpox. The result was a complete breakdown of native life and a precipitous decrease in the Indian population; survivors fled into increasingly inaccessible sections of the Amazon basin. As late as 1906 there were reports of the wholesale capture of Indians who were enslaved in order to tap rubber, which was plentiful and commanded a high price on the world market but which was difficult to exploit because rubber trees were sparsely scattered over a huge area.

Decline of Indian culture

Settlement by Europeans and mestizos (those of mixed Indian and European ancestry) did not occur to any appreciable degree until the 1870s and '80s, when victims of severe droughts in northeastern Brazil began to move into Amazonia to profit from the rubber boom. Another wave

of immigration began at the end of World War II, spurred by the rapid economic development of the region.

Modern settlement patterns. Its vast area notwithstanding, the Amazon basin, in the late 20th century, has a predominantly urban population. Almost one-third of the estimated nine million Brazilians living in the 1.9 million-square-mile area officially designated as Legal Amazonia are concentrated in Belém and Manaus, cities with more than one million inhabitants, and in Santarém. These cities, which are logistic bases of operations for cattle ranching, mining, timber, and agroforestry projects, are still growing rapidly, with modern residential towers and shantytowns standing side by side. Even frontier trading centres in the interior such as Marabá, Pôrto Velho, and Rio Branco have 100,000 or more inhabitants. In the upper reaches of the drainage area, places such as Florencia in Colombia, Iquitos and Pucallpa in Peru, and Santa Cruz in Bolivia have become significant urban centres with most of the amenities of modern life. Air service effectively connects them with Andean and coastal metropolises and even with the more isolated settlements and mission stations of the Oriente.

The economy. *Development of the Amazon basin.* Since World War II the economic development of the Amazon basin has been high on the agenda of every country of which it is a part. From the mid-1940s onward, a number of "penetration roads" have been built from the populous highlands of Colombia, Ecuador, Peru, and Bolivia into the Oriente, which have funneled untold numbers of landless peasants into the lowlands. They also have served to facilitate development of major oil discoveries and timber resources. Tropical hardwoods, river fish, and, since the 1980s, clandestinely produced cocaine have been objects of commercial exploitation, along with Brahman-type livestock raised on pastures newly carved from the selva. Such activities have led to widespread displacement of indigenous groups, who were either forced onto new reserves or left to survive as best they could.

The opening of the Amazon basin has been pursued most aggressively in Brazil. In the mid-1950s the decision was made to refocus the country toward its interior by constructing a new inland capital, Brasília. One consequence of this decision was the initiation of a massive road-building program that aimed at integrating the North (consisting of the states of Amazonas, Acre, and Pará and the territories of Rondônia, Roraima, and Amapá) with the rest of Brazil while establishing an escape valve for the crowded and drought-stricken Northeast. A 1,100-mile-long highway linking Brasília with Belém, the trade centre at the mouth of the Amazon, was completed in 1964. Along with the even more ambitious 3,400-mile all-weather Transamazonian Highway from the Atlantic port of Recife to Cruzeiro do Sul on the Peruvian border—with extensions north to Santarém and Manaus (later to the Venezuelan border) and southward to Cuiabá (Mato Grosso) and Pôrto Velho (Rondônia)—it was to provide the frame for a network of nearly 20,000 miles of highways and feeder routes that was to supersede the traditional fluvial transport system.

Brazilian road-building program

The government had planned to settle about 100,000 families along the Transamazonian Highway, but this goal was not reached. Indeed, the majority of families who did come abandoned the *agrovilas* within a few years because of declining crop yields on the poor soils, weed invasions, plant diseases, lack of credit, and the distance to markets. Disillusioned by the Transamazonian experience, the government shifted its emphasis to encouraging large-scale, capitalist enterprises. Cheap credit and tax breaks were offered to promote the creation of big cattle ranches within Legal Amazonia.

The completion of the Cuiabá–Pôrto Velho highway about 1970 facilitated movement between Mato Grosso and the Rondônia area along the Bolivian border with its more fertile *terra roxa* soils. It brought an unanticipated flood of immigrants from South Brazil, who had become displaced by the shift to large-scale commercial production of export crops (soybeans, citrus, cotton, and wheat). Between 1970 and 1990 the population of Rondônia increased from roughly 116,000 to more than 1,000,000, and that of Acre to the west reached 400,000 by 1990.

Agriculture and forestry. Upland rice, manioc (cassava), and, to a lesser extent, corn (maize) form the mainstay of smallholder agriculture, providing the carbohydrates for the *caboclo* diet. Jute, heart of palm (from *Euterpe oleracea*), and guarana (*Paullinia cupana,* for a favourite Brazilian soft drink) are all minor commercial crops. Black pepper, introduced from Southeast Asia, has become a specialty crop of Japanese colonists.

Cattle pasture by far dominates land use on the cleared parts of the Amazon basin, both in areas of large ranches, such as southern Pará and Mato Grosso, and in areas initially cut over by smallholders for annual crops, as along the Transamazonian Highway. Pasture is even dominant in areas such as Rondônia, where government programs have promoted the cultivation of cacao, coffee, Brazil nuts, and other perennial crops for which a ready cash market exists.

Cattle pasture

Excellent timber is furnished by the mahogany (*Swietenia macrophylla* and *Swietenia humilis*), the Amazonian cedar (*Cedrela odorata*), the Brazilian rosewood (*Dalbergia nigra*), and many other species. Some types, however, are threatened by intensive exploitation. Other trees, such as the coumarou, or tonka bean (*Dipteryx odorata*), yield perfumes, flavourings, and pharmaceutical ingredients. The economic kings of trees, however, are the rubber tree and the Brazil nut. The rubber tree has been one of the most important objectives in the penetration and exploitation of the forest. It gave rise to a period of great but temporary prosperity, especially for the city of Manaus. The rubber gathered from both wild trees and those grown in small plantations continues to make a contribution to the Amazonian economy.

In Brazil areas within the remaining undisturbed forest have been designated for the use of rubber tappers and nut collectors. Yet the establishment of such "extractive reserve" lands has come into conflict with the claims of both squatters and speculators. The latter often have obtained titles by devious means, and their activities require close monitoring.

Corporate farming and agroforestry operations such as Fordlandia, Belterra, and Jari in eastern Brazil and Tournavista in Peru have had little success; the Jari enterprise, for example, was taken over by a consortium of Brazilian investors and the government in 1982. Transnational corporations investing in livestock operations, especially in southern Pará and Mato Grosso, included Volkswagen AG, Swift-Eckrich, Inc., King Ranch, Inc., and Liquigas Italiana. All have terminated their activities.

Mining and energy. The exploitation of the enormously rich mineral complex of the Serra dos Carajás area west of the boom town of Marabá (population 153,000 in 1991) on the Tocantins River has been highly profitable, but it has also had harmful effects on the environment. The site of one of the world's largest and richest iron ore deposits, the district also produces gold, copper, nickel, manganese, tin, and bauxite. The million-acre concession is run by the Companhia Vale do Rio Doce (CVRD), a partnership between private capital and the federal government. Plans for the local smelting of the iron ore could require the clearing of 490,000 acres (200,000 hectares) of forest annually to provide charcoal for producing pig iron. A rail line connects the Carajás development with the Atlantic coast.

Gold mining reached a feverish pitch in the 1980s, stimulated by high world prices of gold. At the height of the Amazon "gold rush," as many as a half million transient miners (*garimpeireos*) came equipped with picks, shovels, and sluice boxes to search for the mineral in the alluvial deposits of the Tocantins valley at Serra Pelada. Brazil's annual production peaked in 1987 at nearly 90 tons, declining thereafter. The mercury used in extracting the gold polluted waterways, causing the fish that are so important in the local diet to become inedible. On the Madeira River, teams operating from rafts pump up from the riverbed auriferous sediments, which have to be subjected to a similar treatment. Bauxite mining, both at Carajás and on the Trombetas River north of the Amazon, requires the use of large settling ponds to trap effluents.

Gold mining

The energy requirements of both the Carajás development and the city of Belém are met by the giant Tucuruí hydroelectric plant on the Tocantins (with a planned power capacity of 7,260 megawatts), the fifth-largest hydroelectric power station in the world. A more modest hydroelectric facility on a small river north of Manaus supplies that city with power. A growing sensitivity to the harmful consequences for both human beings and the environment of the construction of large dams has caused several ambitious projects to be placed on hold.

The principal oil developments within Amazonia have taken place in the Cordillera Oriental of the Andean countries. Oil pipelines lead from producing districts in both Colombia (the upper Putumayo) and Ecuador (Lago Agrio), as well as northeastern Peru, to export terminals on the Pacific coast. Within the Brazilian and Bolivian portions of the basin, developments have been of minimal consequence.

Ecological concerns. International concern about the ecological consequences of continuing deforestation has been growing and was underscored by the United Nations Conference on the Environment and Development ("Earth Summit") held in Rio de Janeiro in 1992. International calls for conservation are based on the view that the Amazon basin is a global resource, one that serves as a control mechanism for the world's climate and as a genetic repository for the future. The nations of the region, however, tend to look upon such calls as a challenge to "national sovereignty."

The extent and rate of deforestation have been subject to continuing controversy. The difficulty of distinguishing via satellite imagery between regenerating secondary vegetation and undisturbed forest as well as the persistence of cloud cover and sometimes smoke have frustrated investigators. The employment of radar has made investigations more precise. It has been suggested that by 1990 some 10 percent of the Amazon selva may have been cleared for pasture, crops, lumber, and firewood. In Brazil deforestation was initiated in Mato Grosso and southern Pará in the 1960s and became widespread over the next two decades in Rondônia and Acre. Already in 1988 Rondônia was estimated to have been deforested by 17 percent, and the process is continuing. In Colombia the upper Putumayo and Caquetá river areas, in Ecuador the province of Napo, and in Peru the Tingo Maria–Pucallpa district have been among the more notable foci of clearing. The cultivation of coca for illicit production of cocaine continues to stimulate such activities.

The consequences of continuing deforestation have been much discussed. Although the forest is an efficient absorber of carbon dioxide, scientists believe that the volume of gas released when substantial parts of the forest are cleared and burned may contribute to global warming through the greenhouse effect. (For further details, see the article HYDROSPHERE, THE: *Impact of human activities on the hydrosphere: Buildup of greenhouse gases.*) Continued conversion of tropical forest to cropland, pasture, or second-growth forest (*capoeira*) may reduce the region's evapotranspiration, thereby interrupting the hydrologic cycle and the recycling of soil nutrients; a likely consequence is an increase in the amount of water running off the surface and greater extremes in water levels.

The unique gene pool of the Amazon Rain Forest, with perhaps two-thirds of the known organisms of the world, is threatened by continuing deforestation. Particular emphasis has been placed on the threat to biodiversity and the possible loss of as yet unknown and unexploited pharmaceuticals contained in the forest.

Finally, at stake is also the survival of many indigenous peoples who, through long residence, have become integrated into the ecosystem of the rain forest and have learned some of its many secrets.

Study and exploration. At the outset of the 19th century, the German explorer Alexander von Humboldt mapped the connection between the Amazon and Orinoco systems through the Casiquiare River. The English naturalist H.W. Bates spent the years from 1848 to 1859 along the Amazon, collecting thousands of species of animals and recording his notes of animals, local peoples, and natural phenomena in a charmingly objective manner. His book, *The Naturalist on the River Amazons,* originally published in two volumes in 1863, is still regarded as one of the great classics on the Amazon River. An official expedition was sent from the United States to Amazonia in the mid-19th century; in 1854 in Washington, D.C., William Lewis Herndon published as a public document the report that he and Lardner Gibbon—both lieutenants in the U.S. Navy—had made to Congress under the title of *Exploration of the Valley of the Amazon.*

The 20th century. The period since 1900 has been one of numerous exploratory and scientific expeditions. In 1913–14, the former U.S. president Theodore Roosevelt and Brazilian Colonel Cândido Rondon headed an expedition that explored a tributary of the Madeira and made natural history collections and observations. A party sponsored by Harvard University's Institute of Geographical Exploration did important scientific work in the years 1910–24. The American Geographical Society compiled data for and published detailed maps of this vast region.

Since World War II, the international scientific community has been increasingly attracted to Amazonia. British, French, German, Japanese, and North American groups have carried out detailed biophysical and cultural surveys; a large number of international workshops, conferences, and symposia on Amazonian problems have been held. Brazilian scientists have also contributed significant research on issues concerning the area. Particularly important has been the work of the National Institute of

International concern

Boat traffic on the Amazon River near Gurupá, Pará state, in Brazil.

Amazonian Research (INPA) at Manaus and the Goeldi Museum in Belém. (R.E.Cr./A.R.S./Ja.J.P.)

ORINOCO RIVER BASIN

The Orinoco River (Spanish: Río Orinoco) and its tributaries constitute the northernmost of South America's four major river systems. Bordered by the Andes Mountains to the west and the north, the Guiana Highlands to the east, and the Amazon watershed to the south, the river basin covers an area of about 366,000 square miles (948,000 square kilometres). It encompasses approximately four-fifths of Venezuela and one-fourth of Colombia. The Orinoco River itself flows in a giant arc for some 1,700 miles (2,740 kilometres) from its source in the Guiana Highlands to its mouth on the Atlantic Ocean. Throughout most of its course it flows through Venezuela, except for a section where it forms part of the frontier between Venezuela and Colombia. The name Orinoco is derived from Guarauno words meaning "a place to paddle"—*i.e.*, a navigable place.

For most of its length, the Orinoco flows through impenetrable rain forest or through the vast grassland (savanna) region of the Llanos (Spanish: "Plains"), which occupies three-fifths of the Orinoco basin north of the Guaviare River and west of the lower Orinoco River and the Guiana Highlands. The savanna was given its name by the Spaniards in the 16th century and long has been used as a vast cattle range. Since the 1930s this region has been developing into one of the most industrialized areas of South America.

Physical features. *Physiography of the Orinoco.* The western slopes of the Sierra Parima, which form part of the boundary between Venezuela and Brazil, are drained by spring-fed streams that give rise to the Orinoco River. The source is placed in Venezuela at the southern end of the Sierra Parima, near Mount Delgado Chalbaud at an elevation of some 3,300 feet (1,000 metres). From its headwaters the river flows west-northwest, leaving the mountains to meander through the level plains of the Llanos. The volume of the river increases as it receives numerous mountain tributaries, including the Mavaca River on the left bank and the Manaviche, Ocamo, Padamo, and Cunucunuma rivers on the right.

The Casiquiare River

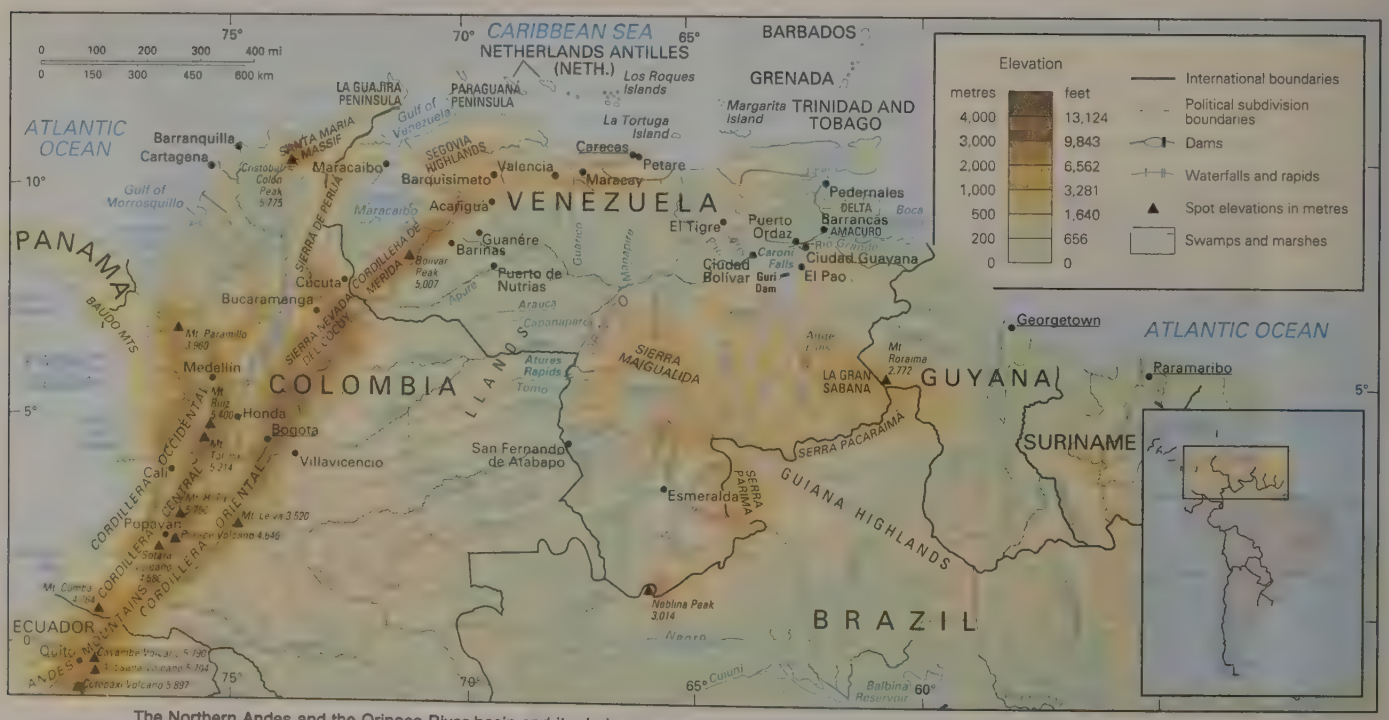
Below the town of Esmeralda, some of the waters of the Orinoco flow south into the Casiquiare River (Brazo Casiquiare; sometimes called the Casiquiare Channel). This channel, a feature peculiar to the Orinoco River system, is a natural passage that flows generally south until it combines with the Guainía River to form the Negro River, thus linking the Orinoco and Amazon river systems.

After its bifurcation in the Casiquiare, the Orinoco bends to the northwest and flows in great meandering curves to its confluence with the Ventuari River. There the river turns to the west to run between high alluvial banks, its course marked by extensive sandbars. Near San Fernando de Atabapo, the Atabapo and Guaviare rivers join the Orinoco, marking the end of the upper Orinoco.

Downstream from San Fernando de Atabapo, the river flows northward and forms part of the border between Venezuela and Colombia. It passes through a transitional zone, the Region of the Rapids (Región de los Raudales), where the Orinoco forces its way through a series of narrow passages among enormous granite boulders. The waters fall in a succession of rapids, ending with the Atures Rapids. In this region, the main tributaries are the Vichada and Tomo rivers from the Colombian Llanos, and the Guayapo, Sipapo, Autana, and Cuao rivers from the Guiana Highlands.

The Atures Rapids mark the beginning of the lower Orinoco basin, in which the river makes its great bend to the east. In this section, the river flows slowly through the lowest level of the plains and increases to about five miles in width. Along the bend, it receives the largest number of tributaries of its entire course, including the Meta, Arauca, and Capanaparo rivers. The Apure River contributes waters from numerous Andean streams, which form a swampy maze in their lower courses.

From its junction with the Apure, the Orinoco meanders eastward over gently sloping plains. Shoals and alluvial islands are abundant; some of the islands are large enough to divide the channel into narrow passages. Tributaries include the Guárico, Manapire, Suatá (Zuata), Pao, and Caris rivers, which enter on the left bank, and the Cuchivero and Caura rivers, which join the main stream on the right. So much sediment is carried by these rivers that they often form islands at their mouths. The Caroní River, one of the Orinoco's largest tributaries, joins the river on its right bank after passing through the Guri Reservoir formed by Guri (Raúl Leoni) Dam, above Ciudad Guayana (also called Santo Tomé de Guayana); farther upstream, on the Churún River (a tributary of the Caroní), are the Angel Falls, the highest waterfall in the world (3,212 feet). Many lagoons, including the Mamo, Amana, and Colorada, are located on the banks of the



The Northern Andes and the Orinoco River basin and its drainage network.

Orinoco River near Ciudad Guayana, in Venezuela.

Orinoco west of its confluence with the Caroní and east of Ciudad Bolívar.

The delta region

About 30 miles downstream of Ciudad Guayana, at the town of Barrancas, the Orinoco begins to form its great delta. The delta extends for about 275 miles along the Atlantic coast, from Pedernales on the Gulf of Paria in the northwest to Barima Point in the southeast on the Boca Grande (literally, the "Great Mouth"). Scores of islands are connected by innumerable canals (*caños*), which form an intricate network. The main channel of the Orinoco, known in the delta as the Río Grande, flows eastward from Barrancas to discharge into the Boca Grande.

Physiography of the Llanos. The Llanos encompasses nearly all of the western lower Orinoco basin, occupying some 220,000 square miles; most of the land is less than 1,000 feet above sea level. The High Plains (Llanos Altos) are most conspicuous near the Andes, where they form extensive platforms between rivers and are some 100 to 200 feet above the valley floors. Away from the mountains they are increasingly fragmented, as in the dissected tableland of the central and eastern Venezuelan Llanos (the Sabana de Mesas) and the hill country (*serranía*) south of the Meta River in Colombia. The Low Plains (Llanos Bajos) are defined by two rivers, the Apure in the north and the Meta in the south. The lowest portion of the Llanos is an area that lies to the west of the lower Orinoco valley; this area is converted annually into an inland lake by flooding.

In addition to the Apure and the Meta, the principal streams draining the Llanos include the Guaviare and Arauca rivers. Seasonal changes between saturation and dehydration have led to advanced laterization of the soil, the process in which the base minerals have been leached away or incorporated into insoluble iron and aluminum silicates. Fine-grained soils form hardpans (cemented layers of soils), and in gravel regions, iron-cemented quartz conglomerates underlie the surface. Excessive acidity and the lack of nutrient bases, organic matter, and nitrogen make virtually all mature soils infertile.

Climate. The climate of the Orinoco basin is tropical, with the seasons marked by differences in rainfall rather than in temperature. The year is divided into two seasons—rainy and dry (locally known as winter and summer)—the former extending from April to October or November and the latter most marked from November through March or April. The wet and dry seasons result from the annual migration of the intertropical convergence zone, a low-pressure trough between the hemispheric easterlies, or trade winds; the passage of the zone northward from its summertime position south of the Equator brings the rainy winter period.

Rainfall varies considerably throughout the drainage basin. The northeast trade winds blow across the coastal districts without losing much of their moisture, in some places leaving less than 20 inches (510 millimetres) of precipitation per year. Areas lying behind topographic barriers also get little rain, while windward slopes generally are well watered. In some regions enough rain falls to support a lush jungle growth, and in others there is enough for a true rain forest (selva). The Llanos experience severe drought from about January to April and then undergo extensive flooding from June to October. Monthly precipitation is seldom less than 10 inches in the Colombian Llanos between April and November. The rains peak about midyear in the Venezuelan north, with monthly totals of roughly 10 inches. Annual precipitation is highest near the Andes, where Villavicencio, Colom., receives 180 inches; and there is a pronounced decrease toward the central plains, where Puerto de Nutrias, Venez., receives 45 inches.

Temperatures

In contrast to precipitation, temperature differences in the basin are slight throughout the year; and no month averages more than 69° F (21° C) or less than 64° F (18° C). Whatever the average temperature, there is little difference from month to month. The only marked variation is from day to night, being greater than that from month to month. On the Llanos, daily maximum temperatures rise above 95° F (35° C) in the dry period; the dry winds and nocturnal cooling bring relief with normal minimum temperatures between 65° and 75° F (18° and 24° C).

Hydrology. The river basin, as a geomorphological feature, dates from the Quaternary Period (*i.e.,* the past 1.6 million years). The enormous quantities of material produced by the highland regions are carried down by torrential rains to the rivers. The rivers, unable to hold the excessive material, overflow or break their banks, producing periodic floods that submerge the lowlands. Under these conditions, drainage presents an unstable and indefinite pattern, marked by the shifting of rivers, lagoons, and swamps over the lower lands. The Orinoco delta is rapidly extending into the ocean, but the tremendous amounts of sediments that accumulate are accelerating the subsidence (sinking) that also is occurring in the delta region.

Wide fluctuations in the river's flow reflect the seasonal rainfall pattern. During the dry season, or "low-water" period, from October to March, the average depth of the Orinoco is about 49 feet in the lower basin near Ciudad Bolívar. The rise of the river begins with manifest regularity in April at the beginning of the rainy season. The "high-water" period from April to October reaches its maximum in July. The depth of the river at this period is about 165 feet at Ciudad Bolívar. From June to August the lowlands of the basin are flooded and in some places are 65 feet under water. At the end of August the waters gradually recede until they again reach their lowest point in October.

Plant life. Most of the Llanos consists of treeless savanna. In the low-lying areas, swamp grasses and sedges are to be found, as is bunchgrass (*Trachypogon*). Long-stemmed grass dominates the dry savanna and is mixed with carpet grass (*Axonopus affinis*), the only natural grass to provide green forage during the dry season.

The most conspicuous trees in the Llanos occur in the gallery forests that occur in the alluvial soils deposited along the rivers and in the narrower files of trees known as *morichales,* named for the dominant moriche, or miriti, palm (*Mauritia flexuosa*), that follow minor water courses. Broad-leaved evergreens originally occupied the high-rainfall zone in the Andean piedmont. There also is a handful of xerophytic trees (*i.e.,* those adapted to arid conditions), including the *chaparro* (scrub oak) and the dwarf palm, scattered on the open savanna. Much of this natural tree

cover, however, has been reduced by deforestation. The Guiana Highlands are covered with high, dense forest that is interrupted by small patches of savanna. The tropical rain forest of the upper Orinoco valley contains hundreds of species of trees. Mangrove swamps cover much of the delta region.

Birds

Animal life. More than 1,000 species of birds frequent the Orinoco region; among the more spectacular are the scarlet ibis, the bellbird, the umbrella bird, and numerous parrots. The great variety of fish include the carnivorous piranha, the electric eel, and the *laulao*, a catfish that often attains a weight of more than 200 pounds. The Orinoco crocodile is the longest of its kind in the world, reaching a length of more than 20 feet; among other inhabitants of the rivers are caimans (a crocodile-like amphibian) and snakes, including the boa constrictor. The arrau, or side-necked turtle, the shell of which grows to a length of about 30 inches, nests on the sandy islands of the river. Insects include butterflies, beetles, ants, and mound-building termites.

The Llanos have few indigenous animals. Most mammals nest in the gallery forests along the streams and feed on the grassland. The only true savanna dwellers in the region are a few burrowing rodents and about two dozen species of birds (among them the white and scarlet ibis, the *morichal* oriole, and the burrowing owl). Several species of deer and rabbit, the anteater and armadillo, the tapir, the jaguar, and the largest living rodent, the capybara, also can be found.

The people. *Indigenous peoples of the basin.* Except for the Guajiros of Lake Maracaibo, most of the Venezuelan aboriginal population lives within the Orinoco River basin. The most important indigenous groups include the Guaica (Waica), also known as the Guaharibo, and the Maquiritare (Makiritare) of the southern uplands, the Warrau (Warao) of the delta region, and the Guahibo and the Yaruro of the western Llanos. These peoples live in intimate relationship with the rivers of the basin, using them as a source of food as well as for purposes of communication.

Settlement. Until the mid-1900s, settlement was limited to widely scattered ranches known as *hatos* ("cow herds"), a few villages, and missionary stations along the lower courses of the region's rivers. Oil strikes in the eastern and central Venezuelan Llanos at El Tigre (1937) and Barinas (1948) initiated industrial and urban development in a region that had been sparsely populated until that time. Several of the "boom towns" of that period, such as El Tigre, have grown into sizable cities. An expansion of intensive agriculture occurred with the settlement, which began in the 1950s, of pioneer farmers in the Andean piedmont and along the river valleys. Major concentrations of these small farms are located in the vicinity of Barinas, Guanare, and Acarigua in Venezuela and in the Ariari region in Colombia.

As a result of this settlement, a high degree of urbanization has occurred in the Venezuelan Llanos, with more than half of the people living in cities of 10,000 or more inhabitants. The important towns, with the exception of Ciudad Bolívar, are built on high ground to avoid recurrent flooding. Town plans reflect Spanish influence: streets are arranged in a gridiron pattern with a central plaza. By contrast, population increase has been modest in the Colombian areas of the Llanos and—with the exception of the region around Ciudad Guayana—in the Guiana Highlands.

The economy. *Resource exploitation.* The Guiana Highlands are rich in mineral deposits. Iron ore, containing high concentrations of iron, is mined at Cerro Bolívar and El Pao. Other minerals include deposits of manganese, nickel, vanadium (a metallic element used to form alloys), bauxite, and chrome. There also are deposits of gold and diamonds. Petroleum and natural gas are exploited in the Venezuelan Llanos and the Orinoco delta.

The Venezuelan Llanos long have been one of South America's major livestock-raising areas, with cattle being predominant. In addition, sugarcane, cotton, and rice are grown on a commercial scale on the plains; and coffee has become important on the northwestern and northern highland fringes of the river basin. Land-reclamation and flood-control projects in the delta region have been planned in order to open vast agricultural lands.

Although agriculture and cattle raising have continued as mainstays of the basin's economy, the base has been widened by the exploitation of petroleum and other minerals and by the establishment of certain industries. Industrial development of the river basin is concentrated around Ciudad Guayana and includes the production of steel, aluminum, and paper. The industrial growth of Ciudad Guayana has been made possible by the construction of the Macagua and Guri dams, which have harnessed much of the immense hydroelectric potential of the Caroní River. The power supplied by this vast project is supplemented by natural gas piped from the oil fields north of the Orinoco River.

The Ciudad Guayana complex

Transportation. The Orinoco and its tributaries long have served as vast waterways for the indigenous inhabitants of the Venezuelan interior. Especially during the floods of the rainy season, boats with outboard motors are the only means of communication throughout large areas of the river basin. Large river steamers travel upriver for about 700 miles from the delta to the Atures Rapids. Dredging has allowed large oceangoing vessels to navigate the Orinoco from its mouth to its confluence with the

Horses being watered on the Llanos, in eastern Colombia.

Caroní River—a distance of about 225 miles—in order to tap the iron ore deposits of the Guiana Highlands.

Considerable road construction has been undertaken in the Venezuelan Llanos since World War II. The Llanos and the Guiana region were connected in 1967 with the completion of a mile-long bridge across the Orinoco at Ciudad Bolívar. Earlier, in 1961, the mouth of the Caroní was bridged to connect the new industrial town of Puerto Ordaz with the old Orinoco port of San Félix, thereby creating the urban unit of Ciudad Guayana; Ciudad Guyana subsequently was connected to Caracas by a major highway.

Study and exploration. European exploration of the Orinoco River basin began in the 16th century. A series of expeditions sponsored by the German banking house of Welser of Augsburg penetrated the Llanos southward across the Apure and Meta rivers. From the east, several Spanish expeditions ascended the river from its mouth without much success. In 1531 the Spanish explorer Diego de Ordaz voyaged up the river, and that same year another Spanish explorer, Antonio de Berrio, descended the Casanare and Meta rivers and then descended the Orinoco to its mouth.

In 1744 Jesuit missionaries reached the Casiquiare River. Alexander von Humboldt, the German naturalist, traveled more than 1,700 miles through the Orinoco basin in 1800. By 1860 steamships were navigating the Orinoco. The source of the river remained in dispute, however, until a Venezuelan expedition finally identified it in 1951.

(Me.F.G./Di.B./Ed.)

RÍO DE LA PLATA SYSTEM

The Río de la Plata (literally, "River of Silver"; in English often called the River Plate) is a tapering intrusion of the Atlantic Ocean occurring on the east coast of South America between Uruguay to the north and Argentina to the south. While some geographers regard it as a gulf or as a marginal sea of the Atlantic, and others consider it to be a river, the majority regard it as the estuary of the Paraná and Uruguay rivers (as well as of the Paraguay River, which drains into the Paraná). The Río de la Plata receives waters draining from the basin of these rivers, which covers much of south-central South America; the total area drained is about 1,600,000 square miles (4,144,000 square kilometres), or about one-fourth of the surface of the continent. Montevideo, the capital of Uruguay, is located on the northern shore of the estuary, and Buenos Aires, the capital of Argentina, is on the southwestern shore.

Rivers of the system

The delta of the Paraná and the mouth of the Uruguay meet at the head of the Río de la Plata. The breadth of the estuary increases from the head seaward, a distance of about 180 miles (290 kilometres): it is 31 miles from the city of Punta Lara on the southern (Argentine) shore to the port of Colonia del Sacramento on the northern (Uruguayan) shore, and 136 miles from shore to shore at the Atlantic extremity of the estuary. To those who regard the Río de la Plata as a river, it is the widest in the world, with a total area of about 13,500 square miles.

Physical features. The Paraná River (Spanish: Río Paraná; Portuguese: Rio Paraná), together with its tributaries, forms the larger of the two river systems that drain into the Río de la Plata. The Paraná—meaning "Father of the Waters" in the Guaraní language—is 3,032 miles (4,880 kilometres) long and extends from the confluence of the Grande and Paranaíba rivers in southern Brazil, running generally southwestward for most of its course, before turning southeastward to drain into the Río de la Plata. The Paraná customarily is divided into two segments: the Alto (Upper) Paraná above the confluence with the Paraguay River and the Paraná proper (or lower Paraná) below the confluence.

Physiography of the Alto Paraná basin. The Grande River rises in the Serra da Mantiqueira, part of the mountainous hinterland of Rio de Janeiro, and flows westward for approximately 680 miles; but its numerous waterfalls—such as the Marimbondo Falls, with a height of 72 feet (22 metres)—makes it of little use for navigation. The Paranaíba, which also has numerous waterfalls, is formed by many affluents, the northernmost headstream being the São Bartolomeu River, which rises just to the east of Brasília.

From its origin in the Grande-Paranaíba confluence to its junction, some 750 miles downstream, with the Paraguay, the Alto Paraná receives many tributaries from both the right and the left. The three most important tributaries—the Tietê, Paranapanema, and Iguaçu rivers—all join the Alto Paraná on its left bank and have their sources within a few miles of the Atlantic coast of Brazil.

The Alto Paraná first flows in a southwesterly direction down a deep cleavage in the southern slope of the ancient Brazilian Highlands, the configuration of which determines its course. Just before it begins to run along the frontier between Brazil to the east and Paraguay to the west, the river has to cut through the Serra de Maracaju (Mbaracuyú), which in the past had the effect of a dam, until the Itaipu hydroelectric dam project was completed there in 1982; the river once expanded its bed into a lake 2.5 miles wide and 4.5 miles long, with Guaíra, Braz., standing on the southern shore. The river's passage through the mountains was, until 1982, marked by the Guairá Falls (Salto das Sete Quedas), which had eight times the water volume of the Niagara River of North America. Since the completion of the Itaipu project's first stage, the falls and lake have been submerged, and a reservoir now extends upstream for some 120 miles and covers more than 700 square miles.

The Iguaçu River

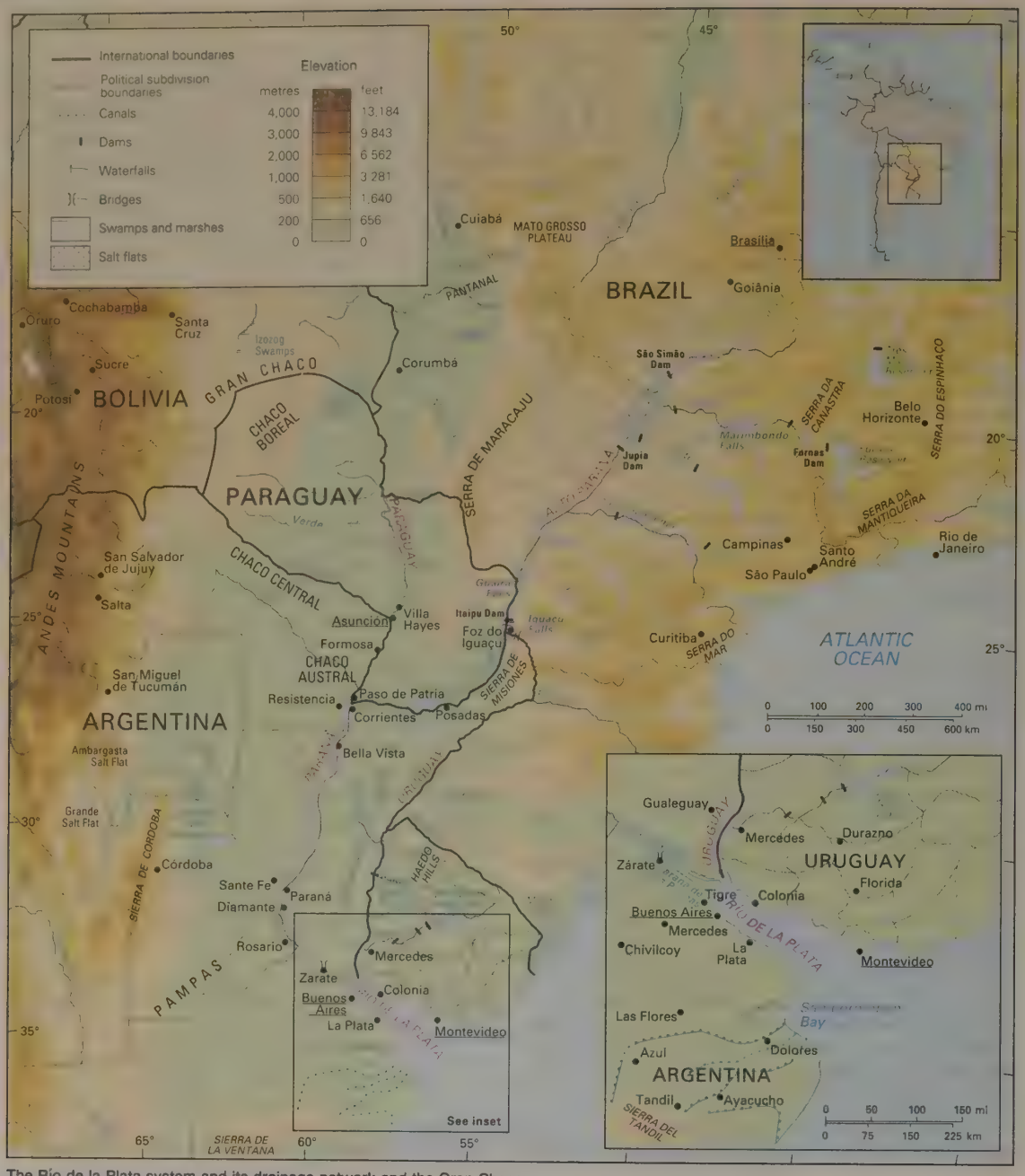
The Iguaçu River (Iguaçu meaning "Great Water" in the Guaraní language) joins the Alto Paraná at the point where Brazil, Paraguay, and Argentina converge. Rising in the Serra do Mar near the Brazilian city of Curitiba (for which reason it is sometimes called the Rio Grande de Curitiba), the Iguaçu flows about 380 miles from east to west, during which some 70 waterfalls reduce the river's elevation by a total of about 2,650 feet. While the Ñacunday Falls are 131 feet high, the spectacular Iguaçu Falls, on the frontier between Brazil and Argentina, 14 miles upstream from the Iguaçu–Alto Paraná confluence, have a height of about 270 feet—almost 100 feet higher than Niagara Falls. As the river approaches the falls, it widens before plunging over the crescent-shaped edge, producing horseshoe-shaped cataracts more than two miles wide. Below the falls, the river passes for several miles through a gorge (Garganta del Diablo; literally, "Devil's Throat") that is only 164 feet wide between heights varying from 65 to 328 feet.

From the Iguaçu confluence to its junction with the Paraguay River, the Alto Paraná continues as the frontier between Paraguay and Argentina. So long as it is flanked on the left (Argentine) bank by the steep edge of the Sierra de Misiones, the river proceeds in a generally southwesterly direction, but it twists repeatedly to and fro over a rocky bed studded with outcrops of porphyritic basalt. At Posadas, Arg., however, where it is about 1.5 miles wide, the river turns abruptly westward and begins a more meandering course, embracing islands of considerable size and punctuated so frequently by rapids and by outcrops of basalt that navigation is difficult. At the Apipé Rapids the river is only about 4 to 6 feet deep.

Physiography of the Paraguay basin. At Paso de Patria, on the right (Paraguayan) bank, the Paraná receives its greatest tributary, the Paraguay River. The fifth largest river in South America, the Paraguay (Spanish: Río Paraguay; Portuguese: Rio Paraguai) is 1,584 miles (2,550 kilometres) long. The name Paraguay, also taken from the Guaraní language, could be translated "river of *paraguas* (coloured, plumed birds)" or "river of cockades," an allusion, perhaps, to the plumed headdresses once worn by the riverine peoples.

The Paraguay also rises in southern Brazil, in the central plateaus of Mato Grosso state, at an altitude of 980 feet above sea level. Where it becomes navigable for small craft—about 150 miles downstream, near Cáceres, Braz., after its confluence with the Sepotuba River—it is 275 feet wide and 20 feet deep. Another 20 miles downstream, where the Jauru River joins it at an elevation of 400 feet, the Paraguay enters the Pantanal, a vast seasonal swamp that covers much of southern Mato Grosso and northwestern Mato Grosso do Sul state. During the dry season

The Pantanal



The Río de la Plata system and its drainage network and the Gran Chaco.

(May to October) the swamps in the Pantanal shrink to small patches of marshy land. With the onset of the rains in November, the slow-flowing rivers are quickly filled to capacity, and a large, shallow lake is formed. Spanish missionaries mistook this for a permanent lake, and it appeared as "Lago Xarays" on early maps of the region.

The Paraguay's main channel skirts the Pantanal's western edge over a sandy bed, flowing around the many islands in its course. During its passage through the Pantanal, the river receives such important tributaries as the Cuiabá, Taquari, and Miranda rivers. About 470 miles downstream, it flows north-south to form the boundary between Brazil and Paraguay before being joined by a tributary, the Apa River, that flows in from the east and demarcates part of the Brazilian-Paraguayan frontier. The river then enters Paraguay, having traveled about 640 miles from its source. After flowing for more than 200 miles across Paraguay, it is joined by the Pilcomayo River at the Argentinian border, near Asunción. It then flows south-southwest along the Argentine-Paraguayan frontier for about 140 miles, until it is joined on its west bank by the Bermejo River. Continuing along the border for another 40 miles, it then empties into the Paraná River at a short distance from the Argentine city of Corrientes.

From its confluence with the Apa for the 630 miles to its mouth, the Paraguay runs on a shallow, broad bed, with an average width of about 2,000 feet. South of Asunción, the river's right (Argentine) bank gradually lowers, whereas its left (Paraguayan) bank becomes elevated, forming cliffs. Along this stretch, floods develop principally on the western bank, spreading over the Argentine plain for distances of from three to six miles. These lands form part of the Gran Chaco.

Physiography of the lower Paraná basin. After its juncture with the Paraguay, the combined stream of the Paraná turns southward as it passes Corrientes. It now becomes a typical "plains" river, banked by its own alluvial deposits and having an extensive floodplain on its right bank, with tracts up to 24 miles wide subject to inundation. Its permanent bed, about 2.5 miles wide at Corrientes, narrows to about 8,000 feet at Bella Vista, to about 7,000 feet at Santa Fe, and to about 6,000 feet at Rosario, and it is strewn throughout with chains of islands. Santa Fe, on the right bank opposite the port of Paraná, stands where the

Paraná receives its last major tributary, the Salado River. Between Santa Fe and Rosario, however, the right bank begins to rise as the river skirts the edge of the undulating plain, which flanks it down to the delta, and reaches altitudes ranging from about 30 to 65 feet. The left bank, meanwhile, is always higher than the right but has to sustain the erosive action of the water, which becomes increasingly turbid as great masses of soil are constantly falling into it; in the delta the main branch of the river runs along a break in the terrain, with its left bank consisting of a cliff about 75 feet high.

The Paraná delta

The delta of the Paraná has its apex as far north as Diamante, upstream from Rosario, where branches of the river begin to turn southeastward. About 11 miles wide at its upper end, the width of the delta grows to roughly 40 miles at the river mouth, where the separated branches of the Paraná flow into the Río de la Plata, about 200 miles from Diamante. With an area of 5,500 square miles, the delta is advancing steadily, as an estimated 165 million tons of alluvial deposits are added annually. Within the delta the river divides again and again into distributary branches, the most important being the two last great channels, the Paraná Guazú and the Paraná de las Palmas. The islands of the delta, alluvial in origin, are low-lying and of varying size. Their shores and the outer fringes of the river have protective embankments covered with trees but nevertheless may be submerged in times of flooding, when they present the appearance of flooded forests.

Physiography of the Uruguay basin. The Uruguay River (Spanish: Río Uruguay; Portuguese: Rio Uruguai) is the other major system, 990 miles (1,593 kilometres) in length, that flows into the Río de la Plata. Like the Alto Paraná and the Paraguay, the Uruguay originates in southern Brazil, formed by several small streams that rise on the western slopes of the Serra do Mar. From the south it is joined by the Pelotas River, which divides the states of Rio Grande do Sul and Santa Catarina. After flowing west, the Uruguay turns southwest at its juncture with the Peperi Guaçu River, the first sizable tributary to join it from the north. For most of its course, the fast-flowing Peperi Guaçu marks the boundary between the Argentine province of Misiones and Brazil; and after its confluence with the Uruguay, the latter river divides Brazil and Argentina. A few miles beyond the juncture with the Peperi Guaçu, the river is constricted between rocky walls in the Grande Falls, a two-mile stretch of rapids with a total descent of 26 feet in 8 miles. At the cataracts, the river narrows suddenly from 1,500 feet to an extreme of 100 feet.

Several small rivers join the Uruguay from the west and are navigable in their lower reaches by canoes and small boats. The principal ones, from north to south, are the Aguapey, Miriñay, Mocoretá (which divides Entre Ríos and Corrientes), and Gualeguaychú. The important tributaries of the Uruguay, however, come from the east. The Ijuí, Ibicuí, and the Cuareim are short rivers but of considerable volume; the last forms part of the boundary between Brazil and Uruguay. At the mouth of the Cuareim, the Uruguay becomes the boundary line between Argentina and Uruguay, and the river flows almost directly south. A dam above the falls at Salto, Uruguay, impounds Lake Salto Grande some 40 miles upstream. The Negro River, approximately 500 miles long and the Uruguay's largest tributary, joins the latter only 60 miles from the Río de la Plata. The Negro rises on the Brazilian border in Rio Grande do Sul state and flows westward through central Uruguay. Like the Alto Paraná, the Uruguay generally is clear and carries little silt, except in the seasonal floods. After its juncture with the Negro, the Uruguay broadens sharply to a width of 4 to 6 miles and becomes a virtual extension of the Río de la Plata estuary.

Physiography of the Río de la Plata. The two contributory river systems bring down an immense quantity of silt each year. The muddiness of the water in the Río de la Plata itself is increased by the tides and winds that hinder the deposition of silt on the bed. When sediments do settle, the mineral and organic matter form great shoals, banks, or bars: the Playa Honda Shoal is just off the Paraná delta, the Ortiz and Chico shoals are farther downstream, and the Rouen, Inglés, Alemán, and Arquímedes shoals are still farther out. The depth of the water—varying from 6 feet above the shoals to 65 feet in the intervening channels—is reduced along the southern coast by an offshore shoal.

Silt and shoals

The Argentine coast of the estuary is low-lying; its banks are of marine debris and coarse sand, and the coast is subject to flooding in places. The entrances to Argentine ports (including that of Buenos Aires) require constant dredging. The Uruguayan coast stands considerably higher and consists largely of rocks and dunes. Off the Uruguayan coast are several small islands, such as Hornos, San Gabriel, López, Lobos, Farallón, and—opposite the mouths of the Uruguay and Paraná Guazú rivers—Martín García.

Hydrology of the system. The velocity of the Paraná's current changes frequently during the river's long course. For the Alto Paraná, the rate becomes slower wherever the bed widens (especially when a real lake is formed, as at Itaipu Dam) and much faster wherever the bed narrows (as in the canyon downstream from Itaipu). Farther downstream, it slackens on its way to Posadas but accelerates thereafter over a series of rapids and races. It becomes slower again downstream from Corrientes, stabilizing its flow at a mean rate of 2.5 miles per hour on the way to the Río de la Plata.

Throughout the basin of the Paraguay River, which covers more than 380,000 square miles, elevations rarely exceed 650 feet above sea level. Thus, over a long distance, the gradient of the river varies only slightly from about 0.75 to 1 inch per mile (1.2 to 1.6 centimetres per kilometre). The various streams of the basin have low banks or natural levees, built up when silt is deposited along the slower-flowing portions of the river channel during flood stage. When the river recedes, its banks thus remain elevated above the level of the neighbouring plains. During floods a continuous water table, often as much as 15 miles wide, underlies the inundated plains, and about 38,600 square miles of surface area are flooded. The Paraguay has varying rates of flow between its source and its confluence with the Paraná. Above Corumbá, in Brazil, it has a typically tropical regime—at its highest in February and at its lowest from July to August. Below Corumbá, the high point occurs in July and the low point from December to January.

The volume of flow

The volume of the lower Paraná is, for practical purposes, correlated to the amount it receives from the Paraguay, which supplies about 25 percent of the total. High periods occur normally between November and February and low periods in August and September. The river's mean overall volume at the Río de la Plata is about 610,700 cubic feet (17,293 cubic metres) per second, with the highest recorded volume being 2,295,000 cubic feet per second (1905) and the lowest 86,400 (1945).

An important factor in the hydrologic regime of the lower Paraná is that the Alto Paraná and the Paraguay reach their maximum flow at different times. Whereas the mountainous basin of the Alto Paraná is drained so rapidly that water begins to rise at Corrientes in November, reaching its maximum height there in February, the Pantanal swamps of the upper basin of the Paraguay retain precipitation so much longer that the Paraguay's high water does not reach Corrientes until May, reaching its maximum in June. Thus, levels on the lower Paraná begin to sink in March, rise from May, and sink again from July to September. Whenever both the Alto Paraná and the Paraguay reach their highest levels at the same time, the lower Paraná has to carry an exceptionally heavy volume of water—as it did in 1905, when the delta experienced heavy flooding.

The volume of water discharged by the Río de la Plata into the Atlantic is estimated at about 776,900 cubic feet per second. Although the water of the tributary rivers is so widely distributed over the length and breadth of the estuary that variations in their volume do not affect the level of the water, the estuary's level is considerably affected by variations of the tides and, especially, of the winds reaching it. The ocean tides are relatively weak, but they flow 120 miles up the Paraná and the Uruguay rivers from their mouths on the estuary. The average tidal range is 0.5 foot at Montevideo and 2.5 feet at Buenos Aires. The pampero

Lush vegetation of the Pantanal, Mato Grosso do Sul state, in Brazil.

(a wind from the south to southwest) and southeasterly winds called *sudestados* both exert a great influence on the Río de la Plata: the pampero, when it is most powerful, drives the water onto the Uruguayan coast, so that the water level drops on the Argentine side; the southeasterly wind has the effect of flooding the Paraná delta and causing the level to drop on the Uruguayan coast.

Climate. The basins of the Alto Paraná and Paraguay have a hot and humid climate throughout the year. The winters (April to September) are dry, and the summers (October to March) are rainy. Annual mean temperatures in the upper basin are above 68° F (20° C), the absolute maximum temperature being from 104° to 107° F (40° to 42° C) and the absolute minimum temperature being about 37° F (3° C). January frequently is the warmest month. More than four-fifths of the annual precipitation occurs in the summer months, with the least amount of rain falling in July and August. Annual rainfall varies from 80 inches (2,000 millimetres) in the mountains to the east to 40 inches in the west. Rainfall takes the form of drenching downpours often accompanied by hailstorms.

The climate of the middle and lower basins progresses from subtropical in the north to temperate in the south. The mean annual temperature along the Río de la Plata is 55° F (13° C), and monthly averages are always over 50° F (10° C). Frosts are frequent in the winter months in the south but can occur as far north as Asunción and Paraná state in Brazil. Humidity in the lower basin is notably high—averaging 70 percent annually along the Río de la Plata—and sometimes is quite stifling in summer; the moist vapours become still thicker when the Paraná brings down the torrential waters of the tropical basin. Rainfall in the southern basin is somewhat less plentiful than in the north, but it occurs at all seasons. The mean annual precipitation along the Río de la Plata is 44 inches.

Plant life. The Brazilian section of the Alto Paraná forms the boundary between two zones: that of the forest to the east and of the savanna to the west. Forests include stands of Paraná pine (*Araucaria angustifolia*), an evergreen conifer valued for its softwood timber. The treeless savanna, with grasses and bushes, is used for cattle raising.

The Pantanal complex

In the upper Paraguay River basin, some of the Pantanal's vegetation, called the "Pantanal complex," is typical of the Mato Grosso Plateau, while the remainder of the basin is typical of lowlands. Plants that thrive in water and in moist soils, as well as those that flourish at moderate temperatures or are adapted to dry regions, are found within the complex. The water plants, found on the permanently flooded lands, are typified by the water hyacinth and by the Amazon, or royal, water lily (*Victoria amazonica*). Moisture-loving species, such as the trumpetwood and the guama, flourish over most of the floodplain. On the savanna, after the floods, various grasses such as paspalum and knotroot bristle grass reappear. Vegetation of a more evolved type, which thrives at moderate temperatures, occupies the unflooded highland. It is represented by nut-bearing palms and by various types of laurels. Dense, evergreen forest galleries grow along stream banks. In the forests of the region, the carandá (a tropical palm that yields a wax similar to carnauba wax), the paratudo, the muriti palm (a large fan palm), and various types of quebracho trees (South American hardwoods that are a source of tannin) predominate.

Farther south, thick, subtropical, semi-deciduous forests extend westward from the Misiones region of Argentina along the Paraná and cover much of eastern Paraguay. These forests provide such decorative hardwoods as lapacho and also contain *Ilex paraguariensis,* a member of the holly family whose roasted leaves are used to prepare the brewed beverage maté. Some forest trees, outside the forest zone proper, still occur in areas of woodland downstream to the Paraná delta. In the Gran Chaco region along the west bank of the river, and in other sections where drought is more pronounced, a thorn forest of xerophytic (drought-tolerant) plants occurs. In the lowlands of eastern Paraguay, forest cover and savanna grasslands alternate.

Animal life. The river system has a rich and varied animal life throughout its length. Among its many edible fish are the dorado (a gold-coloured river fish that resembles a salmon), the *surubí* (a fish with a long rounded body, flattened at the nose), the *patí* (a large, scaleless river fish that frequents deep and muddy waters), the pacu (a large river fish with a flat body, almost as high as it is long), the pejerrey (a marine fish, silver in colour, with two darker bands on each side), and the corbina (white sea bass); the stretch of the Paraná upstream from Corrientes is popular for its dorado sport fishing. Also of note is the meat-eating piranha, a fish resembling the bluegill that travels in large schools and inhabits the tropical parts of the system.

Reptiles include the iguana lizard, two species of caiman (a crocodilian), the water boa, the rattlesnake, the cross viper, and the *yarará* (the most prevalent South American representative of the viper family). Frogs and toads are plentiful, as are freshwater crabs. There are innumerable species of insects and spiders, and the islands are plagued by mosquitos. Herons, cormorants, storks, and game birds also are plentiful.

The people. Before the arrival of the Spanish in the 16th century, the aboriginal population of interior south-central South America was culturally diverse and highly fragmented. The northern basins of the Alto Paraná and Paraguay rivers were inhabited primarily by Guayacurú- and Bororo-speaking peoples. Nomadic hunter-gatherers roamed Mato Grosso and the Pantanal, where the sea-

The original inhabitants

sonally abundant fish were of particular importance. To the south, along the Paraguay and Alto Paraná rivers, the Guaraní occupied semipermanent villages and cleared patches of surrounding forest for the cultivation of corn (maize), cassava (manioc), and other crops. West of the Paraguay River, the Gran Chaco supported sparse populations of nomadic foragers, such as the Lengua and Abipón, as did the Argentine Pampa on the southern shore of the Río de la Plata.

In what is now Paraguay, the Spaniards and Portuguese interbred freely with the indigenous peoples. Consequently, the present riverine population of the country largely is mestizo, or mixed, and Guaraní as well as Spanish is the common language. In Brazil, however, miscegenation was less general, and some groups of indigenous peoples have remained relatively intact, forming isolated nuclei. Others, like the Bororo, Tereno, and Bacairi, constitute minorities who have adopted some aspects of Christianity and Brazilian culture but who also have retained separate tribal identities and live on the fringe of the region. A significant element in the population of the Alto Paraná region of Brazil consists of descendents of mainly German and Japanese immigrants.

The shores of the Río de la Plata now contain the highest population concentrations of the river system and are the most densely populated areas of both Argentina and Uruguay. In contrast to most of the upper basin, this region is populated mainly by people of European descent. Buenos Aires, on the Argentinian shore, is the centre of one of the world's largest urban agglomerations and contains about a third of Argentina's population; Montevideo, on the Uruguayan side, is considerably smaller but still is one of South America's major cities.

The economy. *Resource exploitation.* The economic usefulness of these river systems is not commensurate with the area that they drain. Economic uses to which these rivers might lend themselves, such as irrigation or hydroelectric power, are difficult to achieve. The swamps of the Pantanal and the Chaco long made agriculture a virtual impossibility in these areas. The gradual use of the potential electric power represented by sites such as Itaipu or Foz do Areia, however, has begun to stimulate the development of industry and agriculture.

The economic development of the upper basin has been hindered by limited natural resources. Paraguay and adjacent parts of Brazil and Argentina are virtually devoid of mineral deposits. Industry, therefore, is limited to processing agricultural products, mainly hides and starch from cassava, or gathering such forest products as petitgrain oil from naturally growing citrus trees. Decorative hardwoods from eastern Paraguay and softwood timber (*Araucaria*) have declined in importance as stocks have been depleted. Cattle grazing long has dominated the Chaco, Mato Grosso, and the grasslands of eastern Paraguay, but the difficulty of transporting beef to distant markets has restricted most production to local consumption. Small-scale farming of food crops occupies most of the rural population. Along the Alto Paraná, successful plantations have been established that produce maté, tung oil, and tea. Agriculture

The lower basin also has been a traditional region of livestock production. Corrientes province and the Pampa region near Río de la Plata in Argentina and the prairies of Uruguay long have supported ranches raising high-quality cattle and sheep; livestock products still dominate the exports of both countries. Crops of cotton, flax, and corn are important along the Argentine shore of the Paraná. Uruguay has attempted to diversify its agriculture, but nearly all of the land area is still used for grazing.

Transportation. A major impediment to economic development in the upper part is poor navigation. A large portion of the upper basin cannot be used at all or is limited to vessels with shallow drafts. Elsewhere, navigation can be maintained in many areas only by constant dredging and renovation of port facilities. The value of these river systems as commercial arteries, therefore, is concentrated on the lower reaches.

The current, narrowness, curves, and presence of exposed rock sills on the Alto Paraná restrict considerably the size of vessels, and several rapids can be passed only with the use of winches to pull the vessels. The river's narrowness, whirlpools, and the increased speed of the current make navigation more dangerous as the mouth of the Iguaçu is approached. A railroad to the town of Guaíra circumvents the Itaipu hydroelectric site and opens up another 400 miles of navigable river farther upstream; in addition, a nearby bridge across the river between Brazil and Paraguay and another across the Iguaçu between Brazil and Argentina are vital links in the regional road system. On the Paraguay River, navigation is complicated by the large seasonal fluctuations in the water level. Small oceangoing ships can reach Asunción but risk being stranded during the dry season. Shallow-draft vessels are able to reach Corumbá, Braz., at all seasons, and smaller craft can reach Cáceres.

By contrast, large oceangoing vessels can travel up the lower Paraná as far as Santa Fe or Paraná. Ocean trade also can reach Concepción del Uruguay directly by the Uruguay River. Long fleets of barges carry the bulk of the river freight. For the people living along its shores, the Río de la Plata always has been useful as a waterway. As a thoroughfare for trade, the estuary is important not only to the people of the coasts but also to the inhabitants of the most remote areas of the drainage basin. Buenos Aires is one of the principal seaports of the world and is the main

Bridge over the Alto Paraná River between Ciudad del Este, Paraguay, and Foz do Iguaçu, Braz.

port of Argentina. Both Buenos Aires and Montevideo are concerned primarily with meat and grain exports from the hinterland; the refrigerators, flour mills, and shipyards required for this trade are located in the coastal zone, as are factories for vegetable oils, textile industries, metallurgical plants, and petroleum refineries.

History. *Early European exploration and settlement.* The Río de la Plata was first explored by Europeans in 1516, when an expedition led by Juan Díaz de Solís, chief navigator of Spain, traversed the estuary as part of its effort to find a route to the Pacific; the estuary was temporarily named in memory of Díaz de Solís after his death on its shores at the hands of unfriendly Charrua Indians. The Portuguese navigator Ferdinand Magellan reached the estuary in 1520 and explored it briefly before his expedition continued on its circumnavigation of the globe. Between 1526 and 1529 the Italian explorer Sebastian Cabot made a detailed study of the estuary and explored the Uruguay and Paraná rivers. Cabot ascended the Paraná as far as the present city of Asunción, Paraguay, and also traveled some distance up the Paraguay River; at Asunción he obtained silver trinkets in barter with the Guaraní Indians, and his interest in these objects gave rise to the estuary's permanent name, Río de la Plata, in the hope that it might indeed become a river of silver.

Survey of Sebastian Cabot

Several failed attempts at establishing settlements on the south shore of the estuary (notably near the present location of Buenos Aires) eventually led to explorations upriver and to the founding of Asunción in 1537; Buenos Aires was not refounded until 1580. By about 1610 Jesuit priests had established the first of more than 30 mission settlements that, until the expulsion of the Jesuits in 1767, were the heart of what became known as the "Jesuit Empire." Remarkable ruins of mission churches in Argentina's Misiones province and in eastern Paraguay are all that remain of this extraordinary enterprise. Throughout the Spanish colonial era the Río de la Plata remained a backwash of the empire. The estuary was virtually closed to legal commerce, and Spain ignored the region until Portuguese and English ambitions threatened to expand into the estuary in the 1760s.

Mapping of the basin. The Spaniard Sebastián del Cano, who accompanied the Magellan expedition, was able to include relatively accurate markings of the Paraná, Paraguay, and Uruguay rivers in the map of the estuary that he drew up in 1523. Further cartographic work by agents of the Spanish crown was supplemented considerably by that of Jesuit missionaries, who first covered the entire basin of the Paraná (including the Paraguay River) in an extensive series of maps produced in the 17th century. In the second half of the 18th century, commissioners demarcating the frontiers between Spanish and Portuguese possessions produced a new series of maps. Of later cartographers, the Spanish naturalist and geographer Félix de Azara and the French physician and naturalist Martin de Moussy are the most important.

The region since 1800. Navigation of the river system became a problem when the independent states of Argentina, Uruguay, Brazil, and Bolivia emerged on its courses. Territorial conflicts and restrictions on navigation caused several wars, culminating in the Paraguayan War, or War of the Triple Alliance (1864/65–70), in which Francisco Solano López led Paraguay in a disastrous struggle against Brazil, Uruguay, and Argentina. In the 20th century, similar conflicts, sharpened by rumoured oil wealth, resulted in the Chaco War (1932–35) between Paraguay and Bolivia.

The development of agricultural wealth, particularly in Argentina, resulted in greater appreciation of the commercial value of these river systems after the mid-19th century. Beginning in the 1850s, thousands of German, French, and Italian colonists settled along the lower Paraná River in Santa Fe province. In the 1890s, German pioneers began to carve agricultural settlements from the forests along the Alto Paraná in Paraguay and Argentina. These people later were followed by other Europeans and by a significant number of Japanese.

Wheat, beef, wool, cotton, and hides entered the river and world trade in increasing quantities from Argentina and Uruguay, while from Brazil and Paraguay came forest and tropical products and maté. Port construction and dredging made Buenos Aires more valuable as a seaport, and by 1902 similar improvements had been completed at Rosario. Channel marking, soundings, dredging, and other aids to navigation became a responsibility of all the riparian states. (W.F.O./D.O./N.R.S.)

SÃO FRANCISCO RIVER

One of the great rivers in South America, the São Francisco, 1,811 miles (2,914 kilometres) long, is the fourth largest river system of the continent and the largest river wholly within Brazil. It has been called the "river of national unity" because it long has served as a line of communication between Brazil's maritime and western regions and between the northeast and the south. The river is named for the 16th-century Jesuit leader St. Francis Borgia (São Francisco de Borja). It is an important source of hydroelectric power and irrigation for eastern and northeastern Brazil. The São Francisco basin occupies some 243,700 square miles (631,200 square kilometres).

Physical features. *Physiography.* The São Francisco River rises at about 2,400 feet (730 metres) above sea level on the eastern slope of the Serra da Canastra in southwestern Minas Gerais state, about 150 miles northwest of the city of Belo Horizonte. The river flows for more than 1,000 miles northward across the states of Minas Gerais and Bahia, through the extensive Sobradinho Reservoir, to the twin cities of Juàzeiro and Petrolina. In this stretch the river receives its main left-bank tributaries—the Para-

The upper course

The São Francisco River and its drainage network.

catu, Urucuia, Corrente, and Grande rivers—and its main right-bank tributaries—the Verde Grande, Paramirim, and Jacaré.

About 100 miles below Petrolina, the São Francisco begins a great curve to the northeast and enters a stretch of rapids and falls 300 miles long. In this section the river forms the border between the states of Bahia to the south and Pernambuco to the north. The upper rapids are navigable during periods of high water, but below Petrolina the river is impassable. The broken course—during which the São Francisco receives the São Pedro, Ipueira, and Pajeú rivers—culminates in the great Paulo Afonso Falls. At the top of the falls, the river divides suddenly and violently and cuts three successive falls through the granite rocks for a total drop of about 275 feet. Below the falls the river flows about 190 miles to its relatively narrow mouth on the Atlantic Ocean, about 60 miles northeast of Aracaju. In its lower section the São Francisco is joined by the Moxotó River and forms the border between the states of Sergipe to the south and Alagoas to the north.

Climate and hydrology. The river basin is largely a tropical desert, and the climate generally is hot and dry. The average maximum temperature for the region is 92° F (33° C) and the average minimum 66° F (19° C). The highest temperature recorded is 107° F (42° C). The prevailing winds are from the southeast, east, and northeast. Rainfall is deficient over most of the area, and drought is frequent. Average annual precipitation measures 20 to 40 inches (510 to 1,020 millimetres) in most of the middle basin and 40 to 80 inches in the headwaters region and below the Paulo Afonso Falls; most of the falls zone receives less than 20 inches annually, and a small portion receives less than 10 inches. Precipitation occurs during the summer months (December to March), while the rest of the year—the winter season—is dry.

Because the São Francisco River flows through the driest region of Brazil, it is subject to seasonal changes in water level of up to 30 feet. All its tributaries run dry during the dry season. Until the river was dammed at Juàzeiro, the riverbed upstream from there would vary from a narrow channel during drought periods to a much wider channel during the rainy season; the Sobradinho Reservoir now holds water throughout the year, although its level can vary considerably.

Plant life. The upper (southern), relatively wet part of the basin is covered with savanna (grassland with patches of trees), called *cerrados,* and with forests of mixed evergreen and deciduous trees. Hardwoods include the jacaranda, Brazilian cedar (*cedro*), and vinhatico; cochineal cactus, aloes, and vanilla plants also grow there. Characteristic of the middle river basin is caatinga vegetation (from the Tupí-Guaraní word meaning "white forest"), an area of stunted, often thorny forest. Among the dominant tree species are the leguminous *catingueiras* and *juremas,* members of the euphorbia (spurge) family, and the barriguda tree (a palmlike tree of the Bombacaceae family); bromeliads (plants with basal, often spiny leaves) and a wide variety of cacti constitute the undergrowth. Economically useful species include the caroa, used for its fibres; the castor oil plant; oil, carnauba, and date palms; and cashew and rubber trees.

Caatinga vegetation

The falls zone lies in the dry Brazilian interior, known as the *sertão.* The small amount of rainfall in the area permits the growth of only xerophytic (drought-resistant) brush and grasses. The dry forests of the hilly uplands support carnauba and babassu palms and such plants as the cactus, the rock rose, and the rhododendron. One unusual plant of the *sertão* region is the evergreen *aveloz,* which can grow to heights of 15 to 20 feet and is used as a hedgerow to mark field boundaries. Underground water in the region often is too saline for irrigation or drinking.

The lower São Francisco flows through a floodplain of fine silt soils, and vegetation thrives there. Most of the original tropical semi-deciduous forest that grew along the river, however, has been cleared for agriculture.

Animal life. Animals are not abundant in the São Francisco basin. Among the mammals are various marsupials, rodents (including the large capybara), and skunks; a small wild cat called the *gato de mata;* the titi monkey, which lives in the more wooded areas; and the armadillo, which is ideally suited to the harsh *sertão* environment.

Birds include the partridgelike tinamous, notably the red-winged tinamou (*Rhynchotus rufescens*), as well as doves, parrots, the yellow finch, and the ubiquitous vulture. The drier areas of the basin have some smaller birds, such as hummingbirds and the sabia, a songbird of the thrush family. The birds of the São Francisco basin are greatly prized in Brazil as pets. Their numbers have been reduced, however, by the clearing of habitat for crops and pastures and by hunting.

The river's fish are an important food source. Among the major food species are robalo (a kind of snook), sardines, *pocomó,* and *sarapó.* In addition, the river mouth contains manatees (sea cows) and various species of sharks.

The people and economy. The São Francisco basin in general is inhabited by people of mixed Portuguese, Indian, and African descent. Most people in the coastal region, however, are of predominantly African ancestry, since that region was the main destination of the slave trade to Brazil.

Cassava cultivation along the São Francisco River downstream of Pirapora, Minas Gerais state, in Brazil.

The upper-middle basin is an agricultural region in which cotton, beans, rice, and corn (maize) are grown. The region also produces pineapples, potatoes, maté (tea), melons, sugarcane, coffee, castor and cottonseed oils, and rum. Its major urban centre is Pirapora, Minas Gerais.

Vazante agriculture

The dry *sertão* is used largely for livestock grazing, mainly cattle, goats, sheep, and donkeys. Along the riverbanks *vazante* agriculture is practiced: during the rainy season, shallow waterbeds (*vassantes*) are enclosed by bars of river sediment and support the cultivation of cassava (manioc), corn, beans, and melons. Truck crops are grown on the riverbanks, and carnauba wax, caroa fibre, and rubber are collected. The major city is the market centre of Salvador, the capital of Bahia state. Most of the lower river valley is dry and suitable only for grazing. On the coastal lowlands, rice and sugarcane are grown.

The São Francisco basin contains deposits of agate, gold, iron, diamonds, opals, antimony, galena (the principal ore of lead), mercury, copper, arsenic, manganese, cobalt, and pyrites. There also are deposits of salt, sulfur, alum, marble, limestone, and clay. The river's hydroelectric potential, however, is its most important resource. The main centres of power generation for northeastern Brazil are at the Paulo Afonso Falls and the dam impounding the Três Marias Reservoir. In addition, hydroelectric plants at dams on the upper and lower stretches of the river supply power to the coastal cities of Salvador, Aracaju, and Recife.

The river basin also is important as a source of irrigation water. The largest storage dam, near Petrolina and Juàzeiro, impounds the vast Sobradinho Reservoir. In addition to the Três Marias Reservoir, there is a large reservoir at Pedra d'Água, and there are plans for a reservoir on the Jaguaribe and a 300-mile-long irrigation canal to join the São Francisco at Petrolina to branches in eastern Pernambuco, Paraíba, and eastern Ceará states.

The river is navigable for small steamers for more than 1,000 miles from Pirapora, Minas Gerais, to Petrolina and Juàzeiro. Sandbars at the river's mouth prevent the entry of deep-draft ocean vessels. There is a railway bridge across the river at Pirapora, and another bridge connects the cities of Petrolina and Juàzeiro. Although river transport is slow and difficult, the São Francisco is an important link between the mining districts of Minas Gerais and roads that radiate north and east from Juàzeiro.

Study and exploration. The earliest European forays beyond the coast and into the São Francisco basin consisted of slave raids on the Indian population. Other than some scattered settlement for cattle ranching, little attention was given to the interior until the 19th century. Considerable scientific study was carried out in northeastern Brazil following the malaria epidemic of 1938, and the disease was eradicated from the region by 1940. Modern research has included groundwater inventories, aerial photographic and radar surveys of the basin, and exploration for economically exploitable minerals. In addition, early Indian settlements have been the focus of archaeological excavations, and the region's plants have been studied for their possible pharmaceutical value. (C.E.Ca./K.E.W.)

BIBLIOGRAPHY

General works. Broad geographic surveys of the region and of individual countries include PRESTON E. JAMES and C.W. MINKEL, *Latin America*, 5th ed. (1986); GILBERT J. BUTLAND, *Latin America*, 3rd ed. (1972); HAROLD BLAKEMORE and CLIFFORD T. SMITH, *Latin America*, 2nd ed. (1983); and ARTHUR MORRIS, *South America*, 3rd ed. (1987). JAN KNIPPERS BLACK (ed.), *Latin America: Its Problems and Its Promise*, 2nd ed. (1991), is a collection of authoritative studies in physical, economic, and political geography. Useful annual publications include the *South American Handbook*, a travel guide; and *South America, Central America, and the Caribbean* (irregular), published by Europa Publications. (C.W.M.)

Physical and human geography. *Geology:* CARLOS SCHOBBENHAUS *et al.* (eds.), *Geologia do Brasil* (1984), is a thorough account (in Portuguese) of the geology of the continent's largest country. English-language discussions of the continent's geology are found mostly in periodical literature, such as A. FORERO SUAREZ, "The Basement of the Eastern Cordillera, Colombia: An Allochthonous Terrane in Northwestern South America," *Journal of South American Earth Sciences*, 3(2–3):141–151 (1990); VICTOR A. RAMOS, "The Birth of Southern South America," *American Scientist*, 77(5):444–450 (1989), and "The Tectonics of the Central Andes," in SYDNEY P. CLARK, JR., B. CLARK BURCHFIEL, and JOHN SUPPE (eds.), *Processes in Continental Lithospheric Deformation* (1988), pp. 31–54; and JORGE JULIAN RESTREPO and JEAN FRANCOIS TOUSSAINT, "Terranes and Continental Accretion in the Colombian Andes," *Episodes*, 11(3):189–193 (1988). (V.A.R.)

The land: Descriptions of the physical geography of South America are provided in the general works cited above and in E.J. FITTKAU *et al.* (eds.), *Biogeography and Ecology in South America* (1969); A.C.S. WRIGHT and J. BENNEMA, *The Soil Resources of South America* (1965); JOSÉ A.J. HOFFMAN (compiler), *Atlas climático de America del Sur* (1975), with text in Spanish and English; and such classic accounts as CHARLES-MARIE DE LA CONDAMINE, *A Succinct Abridgement of a Voyage Made Within the Inland Parts of South America*, trans. from French (1747); and ALEXANDER VON HUMBOLDT, *Personal Narrative of Travels to the Equinoctial Regions of the New Continent, During the Years 1799–1804*, trans. from French, 7 vol. (1814–29), available also in later editions. (C.W.M.)

Works on South America's plant and animal life include ADRIAN FORSYTH and KENNETH MIYATA, *Tropical Nature* (1984), an introduction to a number of Neotropical organisms in their forest setting, by two well-traveled naturalists; JEAN DORST, *South America and Central America* (1967), a well-illustrated natural history; JOHN C. KRICHER, *A Neotropical Companion* (1989), an exploration of the tropical rain forest, tropical savannas, and coastal ecosystems; GEORGE GAYLORD SIMPSON, *Splendid Isolation: The Curious History of South American Mammals* (1980), an authoritative introduction to the evolution of South American mammalian fauna that was later complemented by the so-called "Great American Biotic Interchange," in which North American species migrated southward; MICHAEL A. MARES and DAVID J. SCHMIDLY (eds.), *Latin American Mammalogy: History, Biodiversity, and Conservation* (1991), a collection of articles; ROBERT S. RIDGELY and GUY TUDOR, *The Birds of South America*, vol. 1, *The Oscine Passerines* (1989), a discussion of identification, habitat, behaviour, and range of the species of the group; GHILLEAN T. PRANCE (ed.), *Biological Diversification in the Tropics* (1982), a collection of technical papers on the biotic refuges in northern South America that were caused by climatic change; and FRANÇOIS VUILLEUMIER and MAXIMINA MONASTERIO (eds.), *High Altitude Tropical Biogeography* (1986), a series of specialist scientific papers focusing on the adaptive evolution of specific groups of plants, animals, and insects, principally in the Andes. (D.W.Ga.)

The people: A general survey of ethnic origins, geographic distribution of the various groups, and racial relations is found in MAGNUS MORNER, *Race Mixture in the History of Latin America* (1967); and in MARVIN HARRIS, *Patterns of Race in the Americas* (1964, reprinted 1980). NICOLÁS SÁNCHES-ALBORNOZ, *The Population of Latin America* (1974; originally published in Spanish, 1973); and WILLIAM M. DENEVAN (ed.), *The Native Population of the Americas in 1492*, 2nd ed. (1992), are historical summaries. DARCY RIBEIRO, *The Americas and Civilization* (1971; originally published in Portuguese, 1970), studies the cultural complexes of the Americas. RICHARD GRAHAM (ed.), *The Idea of Race in Latin America, 1870–1940* (1990), is a history of the concept and of governmental politics in connection with it. A useful reference source on the subject is ROBERT M. LEVINE, *Race and Ethnic Relations in Latin America and the Caribbean* (1980). The most comprehensive and convenient work on all aspects of information about the South American Indians remains the monumental JULIAN H. STEWARD (ed.), *Handbook of South American Indians*, 7 vol. (1946–59). On indigenous languages, in addition to specific chapters in the previously mentioned works, see ČESTMÍR LOUKOTKA, *Classification of South American Indian Languages*, ed. by JOHANNES WILBERT (1968). JOSEPH H. GREENBERG, *Language in the Americas* (1987), offers a controversial classification that minimizes the number of Indian language families in the Americas. A useful overview is presented in HARRIET E. MANELIS KLEIN and LOUISA R. STARK (eds.), *South American Indian Languages* (1985). Demographic dynamics are studied in WALTER WILLCOX (ed.), *International Migrations*, 2 vol. (1929–31, reprinted 1961), in which the most complete statistics on mass European immigration before 1930 are interpreted; W.D. BORRIE, *The Cultural Integration of Immigrants* (1959); RICHARD W. WILKIE, *Latin American Population and Urbanization Analysis: Maps and Statistics, 1950–1982* (1984); ECONOMIC COMMISSION FOR LATIN AMERICA, *Dynamics and Structure of the Human Settlement Process in Latin America and the Caribbean* (1984), a United Nations study; and DEBRA A. SCHUMANN and WILLIAM L. PARTRIDGE (eds.), *The Human Ecology of Tropical Land Settlement in Latin America* (1989). (Gr.W.K.)

The economy: General surveys of economic conditions, government policies, foreign economic relations, and social devel-

opments include ELIANA CARDOSO and ANN HELWEGE, *Latin America's Economy: Diversity, Trends, and Conflicts* (1992); JOHN SHEAHAN, *Patterns of Development in Latin America: Poverty, Repression, and Economic Strategy* (1987); JAMES DINSMOOR, *Brazil—Responses to the Debt Crisis* (1990); ROBERT N. GWYNNE, *New Horizons? Third World Industrialization in an International Framework* (1990); PEDRO-PABLO KUCZYNSKI, *Latin American Debt* (1988); and RHYS JENKINS, *Transnational Corporations and Uneven Development: The Internationalization of Capital and the Third World* (1987). Two annuals, *Economic and Social Progress in Latin America*, prepared by the Inter-American Development Bank; and *Economic Panorama of Latin America*, prepared by the United Nations' Economic Commission for Latin America and the Caribbean, offer current information. The roles of natural resources, the environment, and agriculture in the economic development of the continent are examined in DAVID GOODMAN and MICHAEL REDCLIFT (eds.), *Environment and Development in Latin America: The Politics of Sustainability* (1991); JOSEPH S. TULCHIN and ANDREW I. RUDMAN (eds.), *Economic Development and Environmental Protection in Latin America* (1991); MICHAEL J. TWOMEY and ANN HELWEGE (eds.), *Modernization and Stagnation: Latin American Agriculture into the 1990s* (1991); JOHN O. BROWDER (ed.), *Fragile Lands of Latin America: Strategies for Sustainable Development* (1989); DENNIS J. MAHAR, *Government Policies and Deforestation in Brazil's Amazon Region* (1989), a brief treatment; and MERILEE S. GRINDLE, *State and Countryside: Development Policy and Agrarian Politics in Latin America* (1986). Special studies of other aspects of continental economy include FELIPE LARRAÍN and MARCELO SELOWSKY (eds.), *The Public Sector and the Latin American Crisis* (1991); ROSEMARY THORP, *Economic Management and Economic Development in Peru and Colombia* (1991); GARY GEREFFI and DONALD L. WYMAN, *Manufacturing Miracles: Paths of Industrialization in Latin America and East Asia* (1990); and LAWRENCE S. GRAHAM and ROBERT H. WILSON, *The Political Economy of Brazil: Public Policies in an Era of Transition* (1990). (E.C.G.)

Geographic features of special interest. *Landforms:* On the Andes, classic works on the geography and geology include ALAN G. OGILVIE, *Geography of the Central Andes* (1922); and ISAIAH BOWMAN, *The Andes of Southern Peru* (1916). The role of plate tectonics in the formation of the Andes is discussed in R.W.R. RUTLAND, "Andean Orogeny and Ocean Floor Spreading," *Nature*, 233(5317):252–255 (1971); and DAVID E. JAMES, "The Evolution of the Andes," *Scientific American*, 229(2):60–69 (August 1973). HAROLD OSBORNE, *Indians of the Andes: Aymaras and Quechuas* (1952, reissued 1973), is still a useful discussion of the indigenous peoples. More recent works are SHOZO MASUDA, IZUMI SHIMADA, and CRAIG MORRIS (eds.), *Andean Ecology and Civilization* (1985), a collection of conference papers; *Lost Crops of the Incas: Little-Known Plants of the Andes with Promise for Worldwide Cultivation* (1989), a report prepared by a panel of the National Research Council; BENJAMIN S. ORLOVE and GLYNN CUSTRED (eds.), *Land and Power in Latin America: Agrarian Economies and Social Processes in the Andes* (1980); and WILLIAM P. MITCHELL, *Peasants on the Edge: Crop, Cult, and Crisis in the Andes* (1991). The classic of Andean exploration by EDWARD WHYMPER, *Travels Amongst the Great Andes of the Equator* (1892), is available in later editions. (N.R.S.)

Literature in English for the Gran Chaco and Patagonia is scarce. Early descriptions of the Gran Chaco include LUIS JORGE FONTANA, *El Gran Chaco* (1881, reprinted 1977), the diary (in Spanish) of an 18th-century explorer; and JOHN GRAHAM KERR, *A Naturalist in the Gran Chaco* (1950, reprinted 1968), an account of the author's expeditions in 1889–97. More recent discussions of the two regions can be found in the relevant sections of HERBERT WILHELMY and WILHELM ROHMEDER, *Die La Plata Länder: Argentinien, Paraguay, Uruguay* (1963); and PRESTON E. JAMES, *Latin America*, 4th ed. (1969); and in PHILIP CARAMAN, *The Lost Paradise: An Account of the Jesuits in Paraguay, 1607–1768* (1975); BRUCE CHATWIN, *In Patagonia* (1977, reprinted 1988); and JOHN RENSHAW, "Property, Resources and Equality Among the Indians of the Paraguyan Chaco," *Man*, 23(2):334–352 (June 1988). (K.E.W.)

Drainage systems: Introductory overviews of the Amazon River basin include HELEN SCHREIDER and FRANK SCHREIDER, *Exploring the Amazon* (1970), a well-illustrated descriptive book; the treatment of the basin in N. MARK COLLINS (ed.), *The Last Rain Forests* (1990), pp. 110–129; CATHERINE CAUFIELD, *In the Rainforest* (1985); and HILGARD O'REILLY STERNBERG, *The Amazon River of Brazil* (1975). The issue of tropical forest conversion and its ecologic impact came to public attention with the appearance of R.J.A. GOODLAND and H.S. IRWIN, *Amazon Jungle: Green Hell to Red Desert?* (1975). Collections of essays on the basin, often elaborating on this theme, include HAROLD SIOLI (ed.), *The Amazon: Limnology and Landscape Ecology of a Mighty Tropical River and Its Basin* (1984); ROBERT E. DICKINSON (ed.), *The Geophysiology of Amazonia* (1987); MARIANNE SCHMINK and CHARLES H. WOOD (eds.), *Frontier Expansion in Amazonia* (1984); and JOHN HEMMING (ed.), *Change in the Amazon Basin*, 2 vol. (1985). SUSANNA HECHT and ALEXANDER COCKBURN, *The Fate of the Forest: Developers, Destroyers, and Defenders of the Amazon* (1989), is an overarching historical survey, richly documented, with a critical examination of the political, social, and economic background of the escalating degradation of the Amazon environment. Other studies of people and society, mostly with emphasis on Brazil, include MARIANNE SCHMINK and CHARLES H. WOOD, *Contested Frontiers in Amazonia* (1992); JULIE SLOAN DENSLOW and CHRISTINE PADOCH, *People of the Tropical Rain Forest* (1988); EUGENE PHILIP PARKER (ed.), *The Amazon Caboclo* (1985); JOHN HEMMING, *Amazon Frontier: The Defeat of the Brazilian Indians* (1987); and, for recent archaeological discoveries, ANNA C. ROOSEVELT, "Secrets of the Forest," *The Sciences*, 32:22–28 (November/December 1992). Works on resources and ecology include ENEAS SALATI *et al.*, "Amazonia," in B.L. TURNER II *et al.*, *The Earth as Transformed by Human Action* (1990), pp. 479–493; DAVID CLEARY, *The Brazilian Rainforest: Politics, Finance, Mining, and the Environment* (1991); KENT H. REDFORD and CHRISTINE PADOCH (eds.), *Conservation of Neotropical Forests* (1992); HILGARD O'REILLY STERNBERG, "Manifest Destiny and the Brazilian Amazon," in CONFERENCE OF LATIN AMERICANIST GEOGRAPHERS, *Yearbook*, vol. 13 (1987), pp. 25–35; MICHAEL GOULDING, *Amazon: The Flooded Forest* (1989); WILLIAM M. DENEVAN and CHRISTINE PADOCH (eds.), *Swidden-Fallow Agroforestry in the Peruvian Amazon* (1988); PHILIP M. FEARNSIDE, *Human Carrying Capacity of the Brazilian Rainforest* (1986); D.A. POSEY and W. BALÉE (eds.), *Resource Management in Amazonia* (1989); and, on dwindling wildlife, NIGEL J.H. SMITH, *Man, Fishes, and the Amazon* (1981); and KENT H. REDFORD, "The Empty Forest," *BioScience*, 42(6):412–422 (June 1992). Broader surveys of the economic development of the basin include *Minimum Conflict: Guidelines for Planning the Use of American Humid Tropic Environments* (1987), prepared by the United Nations Environment Programme in cooperation with the government of Peru; and STEPHEN G. BUNKER, *Underdeveloping the Amazon* (1985), with an overview of economic history. HENRY WALTER BATES, *The Naturalist on the River Amazons*, 2 vol. (1863, reissued 1989); ALFRED RUSSEL WALLACE, *A Narrative of Travels on the Amazon and Rio Negro*, 2nd ed. (1889, reprinted 1972); and RICHARD SPRUCE, *Notes of a Botanist on the Amazon & Andes*, 2 vol. (1908, reissued 1970), are classics of natural history exploration. WM. LEWIS HERNDON and LARDNER GIBBON, *Exploration of the Valley of the Amazon*, 2 vol. (1853–54), is also informative. (Ja.J.P.)

Detailed studies in Spanish of the physical and human geography of the Orinoco River basin include RAFAEL GÓMEZ PICÓN, *Orinoco, río de libertad*, 2nd rev. ed. (1978); and FRANCISCO TAMAYO, *Los Llanos de Venezuela*, 2nd ed. (1987). The most useful English-language description of the basin still is found in the essays by Alexander von Humboldt in his work cited above in the section on the land. Specialized studies include CARL F. NORDIN, JR., and DAVID PEREZ-HERNANDEZ, *Sand Waves, Bars, and Wind-Blown Sands of the Río Orinoco, Venezuela and Colombia* (1989); and EDWIN D. MCKEE, *Sedimentary Structures and Textures of the Río Orinoco Channel Sands, Venezuela and Colombia* (1989). Among the works on the people of the basin are JOHANNES WILBERT and MIGUEL LAYRISSE (eds.), *Demographic and Biological Studies of the Warao Indians* (1980); and JOHANNES WILBERT, "Geography and Telluric Lore of the Orinoco Delta," *Journal of Latin American Lore*, 5(1):129–150 (Summer 1979), also on the Warao Indians. The settlement of the Llanos is described in JANE M. RAUSCH, *A Tropical Plains Frontier: The Llanos of Colombia, 1531–1831* (1984), and *The Llanos Frontier in Colombian History, 1830–1930* (1993). (Ed.)

English-language sources on the other major river basins are scarce. For the Paraná-Paraguay-Rio de la Plata basin, see M.M. COLE, "Cerrado, Caatinga, and Pantanal: The Distribution and Origin of the Savanna Vegetation of Brazil," *The Geographical Journal*, 76(2):168–179 (June 1960); ERNST GRIFFIN, "The Agricultural Land Use Regions in Uruguay," *Revista Geográfica*, 76:121–151 (June 1972); MARK JEFFERSON, *Peopling the Argentine Pampa* (1926, reprinted 1971); ROBERT C. EIDT, *Pioneer Settlement in Northeast Argentina* (1971); and *Paraguay: Regional Development in Eastern Paraguay* (1978), a World Bank country study. (N.R.S.)

The São Francisco River is addressed in KEMPTON E. WEBB, *The Changing Face of Northeast Brazil* (1974); MANUEL CORREIA DE OLIVEIRA ANDRADE, *The Land and People of Northeast Brazil* (1980; originally published in Portuguese, 1973); and JOSÉ AMAURY DE ARAGÃO ARAÚJO *et al.* (eds.), *Dams in the Northeast of Brazil* (1982; originally published in Portuguese, 1982). (K.E.W.)

South Asian Arts

South Asia, consisting of the huge subcontinent of India, includes Pakistan, Bangladesh, Sri Lanka (formerly Ceylon), as well as the nation of India itself. In spite of differences in physical appearance, complexion, stature, and other ethnological features, the people of the entire region of South Asia are unified by a common cultural and ethical outlook; a wealth of ancient textual literature in Sanskrit, Prākrit, and regional languages is a major unifying factor. Music and dance, ritual customs, modes of worship, and literary ideals are similar throughout the subcontinent, even though the region has been divided into kaleidoscopic political patterns through the centuries.

The close interrelationship of the various peoples of South Asia may be traced in their epics, as in the *Rāmāyaṇa* and the *Mahābhārata*. Kinship between the gods and heroes of regions far distant from each other is evident, and the place-names themselves often evoke common sources. Moreover, there have been continual attempts to impose a political unity over the region. In the 3rd century BC, for example, the emperor Aśoka had almost all of this region under his sway; in the 11th century AD, Rājendra I Cōḷa conquered almost the whole of India and a good portion of Southeast Asia; and the great Mughal Akbar again achieved this in the 16th century. Though the expansion and attenuation of boundary lines, the bringing together or pulling apart politically of whole regions, have characterized all of South Asian history, the culture has remained essentially one.

The geography of the region encouraged a common adoration of mountains and rivers. The great Himalayas, which form the northern boundary, are the loftiest of mountains and are conceived to be the embodiment of nobility, the abode of immaculate snow, and the symbol of a cultural ideal. Similarly, the great rivers such as the Brahmaputra and the Indus are regarded as the mothers of their respective regions, assuring prosperity through their perennial supply of water.

The association of lakes and springs with water sprites and sylvan fairies, called *nāga*s and *yakṣa*s, is common throughout the region. Karkoṭa, the name of an early dynasty, itself signifies *nāga* worship in Kashmir. Sculptures of *nāga*s and *yakṣa*s found in widespread sites suggest a common spirit of adoration, as do sculptures, paintings, temples, and religious texts that for centuries were preserved within an oral tradition without losing their immaculate intonation. The same classical dance is seen in sculpture in Gandhāra in Pakistan, in Bhārhut in the north, and in Amarāvatī in the south.

The relation of the various arts to each other is very close in South Asia, where proficiency in several arts is necessary for specialization in any one. Thus, it is believed that without a good knowledge of dance there can be no proficiency in sculpture, for dance, like painting or sculpture, is a depiction of all the world. For its rhythmic movements and exposition of emotion, dance also requires musical accompaniments; hence, knowledge of musical rhythm is essential. For the stirring of emotion either in music or in dance, knowledge of literature and rhetoric is believed to be necessary; the flavour (*rasa*) to be expressed in music, dance, sculpture, or painting requires a literary background. Thus all the arts are closely linked together.

The arts were cultivated in South Asia not only as a noble pastime but also in a spirit of dedication, as an offering to the Almighty. Passages in literature refer to princes studying works of art for possible defects. One inscription that mentions the name of the *sūtra-dhāra* ("architect") of the 8th-century Mallikārjuna temple at Pattadakal epitomizes the accomplishments and ideals, in both theory and practice, of the artist.

Artists traditionally have enjoyed a high position in South Asian societies. Poets, musicians, and dancers held honoured seats in the royal court. An inscription mentions the appreciation bestowed by Rājendra Cōḷa on a talented dancer, and the architect of the temple at Tiruvoṟṟiyūr, who was also patronized by Rājendra, was eulogized for his encyclopaedic knowledge of architecture and art. Nonetheless, the folk arts were closely linked with the elite arts. Tribal group dances, for example, shared common elements with classical art, dance, and music. Among the artistic traditions of the Indian subcontinent, sculpture in the round (*citra*) is considered the highest artistic expression of form, and sculpture in relief (*ardhacitra*) is next in importance. Painting (*citrābhāsa*, literally "the semblance of sculpture") ranks third. Feeling for volume was so great that the effect of chiaroscuro (*i.e.*, use of light and shade to indicate modelling) was considered very important in painting; a passage from a drama of the 5th-century poet Kālidāsa describes how the eye tumbles over the heights and depths suggested in the modelling of a painting. A classical text on art, *Citrasūtra* enumerates noteworthy factors in paintings: the line sketch, firmly and gracefully drawn, is considered the highest element by the masters; shading and depiction of modelling are valued by others; the decorative element appeals to feminine taste; and the splendour of colour appeals to common taste. The use of a minimum of drawing to produce the maximum effect in suggesting form is considered most admirable.

Portraits play an important role in the visual arts of South Asia, and there are many literary references to the effective depiction of portraits both in painting and in sculpture. A 6th-century text, the *Viṣṇudharmottara*, classifies portraiture into natural, lyrical, sophisticated, and mixed, and men and women are classified into types by varieties of hair—long and fine, curling to right, wavy, straight and flowing, curled and abundant; similarly, eyes may be bow-shaped, of the hue of the blue lotus, fishlike, lotus-petal-like, or globular. Artistic stances are enumerated, and principles of foreshortening are explained. Paintings or sculptures were believed to take after their creators, even as a poem reflects the poet.

Although South Asia has continually been subjected to strong outside influences, it has always incorporated them into native forms, resulting not in imitation but in a new synthesis. This may be seen even in the art of the Gandhāra region of Pakistan, which in the 4th century BC was immersed in Greco-Roman tradition. In the sculpture of this period Indian themes and modes have softened the Western style. Foreign influence is evident after the invasion of the Kushāns in the 1st century AD, but the native element predominated and overwhelmed the foreign influence. During the Mughal period, from the 16th century, when Muslims from Central Asia reigned in South Asia, the blend of Iranian and Indian elements produced a predominantly Indian school that spread throughout the region, making it a unified cultural area under imperial rule. The influence of Islāmic art was enhanced by the second Mughal emperor, Humāyūn, who imported painters from the court of the Shāh of Persia and began a tradition that blended Indian and Persian elements to produce an efflorescence of painting and architecture.

Art in all these regions reflects a system of government, a set of moral and ethical attitudes, and social patterns. The desire of kings to serve the people and to take care of them almost as offspring is evident as early as the 3rd century BC. The ideal of the king as the unrivalled bowman, the unifier, the tall and stately noble spirit, the sacrificer for the welfare of the subjects, and the hero of his people (who conceive of him on a stately elephant) is comprehensively illustrated in a magnificent series of coins from the Gupta Empire of North India of the 4th–6th centuries. The concepts of righteous conquest and righteous warfare

are illustrated in sculpture. The long series of sculptures illustrating the history of the South Indian Pallava dynasty of the 4th–9th centuries gives an excellent picture of the various activities of government—such as war and conquests, symbolic horse sacrifices, the king's council, diplomatic receptions, peace negotiations, the building of temples, appreciation of the fine arts (including dance and music), and the coronation of kings—all clearly demonstrating what an orderly government meant to the people. Similarly, moral attitudes are illustrated in sculptures that lay stress on *dharma*—customs or laws governing duty. The doctrine of *ahiṃsā*, or noninjury to others, is often conceived symbolically as a deer, and the ideal of a holy place is represented as a place where the deer roams freely. The joy in giving and renunciation is clearly indicated in art. Sculptures illustrate simple and effective stories, as from the *Pañca-tantra*, one of the oldest books of fables in the world. The spirit of devotion, faith, and respect for moral standards that has throughout the centuries pervaded the subcontinent's social structure is continuously represented in South Asian painting and sculpture.

(C.S.)

The article is divided into the following major sections:

Literature 635
- Sanskrit, Pāli, and Prākrit literatures: 1400BC–AD1200 636
- Dravidian literature: 1st–19th century 641
- Indo-Aryan literatures: 12th–18th century 644
- Islāmic literatures: 11th–19th century 646
- Sinhalese literature: 10th century AD to 19th century 649
- Modern period: 19th and 20th centuries 649

Music 652
- Folk, classical, and popular music 652
  - Rural areas
  - Classical music
  - Nonclassical music of the cities
- Antiquity 653
  - Verdic chant
  - The classical period
- Medieval period 656
  - Precursors of the medieval system
  - Further development of the grāma-rāgas
- The Islāmic period 656
  - Impact on musical genres and aesthetics
  - Theoretical developments
- The modern period 657
  - Rhythmic organization
  - Musical forms and instruments
  - Interaction with Western music

Dance and theatre 660
- The performing arts in India 660
- Indian dance 661
- Indian theatre 667
- Sri Lanka (Ceylon) 670
- Dance and theatre in Kashmir 671
- Pakistan 671
- Bangladesh 673

Visual arts 673
- Visual arts of the Indian subcontinent 673
  - General chracteristics
  - Architecture
  - Sculpture
  - Painting
  - Decorative arts
- Visual arts of Sri Lanka (Ceylon) 706
  - Architecture
  - Sculpture
  - Painting

Bibliography 708

Literature

The peoples of South Asia have had a continuous literature from the first appearance in the Punjab of a branch of the Indo-European-speaking peoples who also settled all of Europe and Iran. In India this branch of Indo-Aryans, as they are usually called, met earlier inhabitants with different languages and no doubt a different culture—possibly a culture akin to that of the Indus Valley civilization, which had a script, and perhaps a literature of its own, of which nothing is known. Certain to have been settled in India were peoples who spoke languages of Dravidian origin, as well as other languages, called Munda, now preserved only by aboriginal tribes, which show affinities with the languages of Southeast Asia.

The earliest literature is of a sacred character and dates from about 1400 BC in the form of the Rigveda. This work stands at the beginning of the literature of the Veda, or canonical Hindu sacred writings, which as a whole is roughly contemporary with the settlement of the Indo-Aryan peoples in the Punjab and farther east, in the mesopotamia of the Ganges and Yamunā rivers. The language of the Rigveda, which is a compilation of hymns to the high gods of the Aryan religion, is complex and archaic. It was simplified and codified in the course of the centuries from 1000 to 500 BC, which saw the development of prose commentaries called the *Brāhmaṇas*, *Āraṇyakas*, and *Upaniṣads*. While there must have been a long tradition of grammarians, the final codification of the language is ascribed to Pāṇini (5th or 6th century BC), whose grammar has remained normative for the correct language ever since. This language is called Sanskrit (Tongue Perfected). Sanskrit has had a scarcely interrupted literature from about 600 BC until today, but its greatest efflorescence was in the classical period, from the 1st to 7th centuries AD. Because it was identified with the Brahminical religion of the Vedas, reform movements such as Buddhism and Jainism disdained the use of Sanskrit and adopted literary languages—amalgams of different dialects of the parent language—of their own, Pāli in Buddhism and Ardhamāgadhī in Jainism. These languages, usually called Prākrits—that is, derivative as well as more "natural" languages—produced a vast and, again, mostly sacred literature. In a further development of these dialects, the early beginnings can be seen of modern Indo-Aryan languages of northern India: Bengali (also the language of Bangladesh), Hindi (the official language of the Republic of India since 1947), Rajasthani, Punjabi, Gujarati, Marathi, Kashmiri, Oriya, Assamese, and Sindhi, each of which produced a literature of its own. Their names are derived from the regions in which they are spoken, regions with uncertain boundaries, where the different dialects fused at the borders. They all retained a close family resemblance that made bilingualism easy and a fact of Indian literary life.

Far more marked was the difference between Indo-Aryan speech and the languages of the Dravidian family, which are structurally wholly different, though in time a measure of convergence took place. Among them, the oldest recorded is Tamil, now the language of Tamil Nadu (Madras) state and of northern Sri Lanka, whose literature goes back to the early centuries of the Christian Era. Later to be put to literary uses were the cognate Telugu (Andhra Pradesh), Kannada (or Kanarese, Mysore state), and Malayalam (Kerala state) languages.

Influence of Sanskrit

In spite of this linguistic differentiation, the literatures composed in all of these languages reflect, in different degrees, the monumental influence of Sanskrit literature, Sanskrit being the universal Indian language of culture. This influence was one of both substance and form: in substance it provided the basic themes of literary enterprise, notably through the epics, the *Mahābhārata* and *Rāmāyaṇa*, the Hindu popular texts of the *Purāṇas*, especially the *Bhāgavata*, and the mythological repertory that came with Sanskritic Hinduism; in form, Sanskrit belles lettres bequeathed models of literary composition, and Sanskrit poetics provided the aesthetic theory underlying the models. The impact of Islām created a new language, Urdu (from Persian: Camp), based on Hindi; Urdu was the lingua franca of the army. Urdu was used later for literature and at present is the mother tongue of most Indian Muslims and their brethren in Pakistan. Its influence, however, does not compare with that of Sanskrit.

Comparable to the impact of Sanskrit, but far more alien, is that of English, which began to assert itself in the 18th century. The language brought with it new literary forms that were gradually adapted to the old ones, producing new genres—without necessarily giving up the older ones—in the local languages and giving rise to an

interesting literature in the English language. Once more, a universal cultural language to a large extent unified aims in the scattered languages; English still plays this role, though it appears to be slowly declining.

SANSKRIT, PĀLI, AND PRĀKRIT LITERATURES: 1400 BC–AD 1200

Sanskrit: formative period (1400–400 BC). The oldest document in the literature of South Asia is the Rigveda, or Veda of the Stanzas (*c.* 1400 BC), the fundamental text of Brahminical Hinduism. Not literary but religious-magical in its purposes, it is mostly a compilation of hymns, dedicated to a number of gods of the Vedic religion. They have the regular structure of an invocation: the attention of the god is evoked; a brief account of some of his feats is given, to hold his attention; and an exhortation for his help concludes the hymn. The poets, of whom little is known, appear to have come at the close of a priestly poetical tradition, rivalling one another in allusions to obscure exploits, in language often opaque and at times intended to mystify. Nevertheless, the Rigvedic hymns include lines of great beauty. They may occur in a riddling verse, such as "When the ancient Dawns first dawned, the great Syllable was born in the footsteps of the Cow," alluding to the birth of speech at the beginning of creation. Or they may occur in poetry addressed to a deity whose beauty inspires the poet to well-turned lines. To the Dawns, for example: "They approach equally in the east, spreading themselves equally from the same place./ The Goddesses waking from the seat of order, like herds of kine set loose, the Dawns are active"; or to the goddess of the night: "Night coming on, the goddess shines/ In many places with her eyes:/ All-glorious she has decked herself./ Immortal goddess, far and wide/ She fills the valleys and the heights:/ Darkness with light she overcomes."

Nonsacred verses are very rare in the Rigveda, but, when they occur, they can be quite powerful, as in a hymn of a gambler, who is speaking:

> It pains the gambler when he sees a woman,
> Another's wife, and their well-ordered household:
> He yokes these brown steeds early in the morning,
> And, when the fire is low, sinks down an outcast.
>
> "Play not with dice, but cultivate thy cornfield;
> Rejoice in thy goods, deeming them abundant:
> There are thy cows, there is thy wife, O gambler."
> This counsel Savitri the kindly gives me.

Influence of the Rigveda on Sanskrit poetry

Although not literary in purpose, the Rigveda had a decisive influence on the form of Sanskrit poetry: except for narrative verse, the basic unit of all subsequent poems (no matter how many verses they consist of) is the single stanza that contains one complete thought.

The second Veda (*c.* 1200 BC), the Yajurveda (Veda of the Yajus [Formulas]), contains sacred formulas recited by a group of priests at the great Vedic sacrifices; and the third (*c.* 1100 BC), the Sāmaveda (Veda of the Chants), is in essence an anthology of the Rigveda. More literary interest attaches to the fourth Veda (1200 BC), the Atharvaveda (an *atharvan* was a special priest), which contains hymns, incantations, and many magic charms.

The succeeding literature (*c.* 1000–700 BC), the *Brāhmaṇas* ("Disquisitions About the Ritual"), continues not the poetry but the liturgical concerns of the Rigveda. They were written in a dry, expository prose, so that only their narrative portions have any literary interest. Much the same is true of the next layer of Vedic texts (800–600 BC), the *Āraṇyakas* ("Books Studied in the Forest").

The *Upaniṣads*

But the picture changes in the *Upaniṣads* (*c.* 1000–500 BC; "Collections of Esoteric Equations"). These prose texts at times convey the actual mode of teaching of a revered sage, in a style that can be strikingly intimate:

> "Bring me a fruit of that *nyagrodha* (banyan) tree."
> "Here it is, venerable Sir."
> "Break it."
> "It is broken, venerable Sir."
> "What do you see there?"
> "These seeds, exceedingly small, venerable Sir."
> "Break one of these, my son."
> "It is broken, venerable Sir."
> "What do you see there?"
> "Nothing at all, venerable Sir."
> The father said: "That subtle essence, my dear, which you do not perceive there—from that very essence this great *nyagrodha* arises. Believe me, my dear."

While the older *Upaniṣads* are in prose, the later ones, dating from around 500 BC, mark a shift back to verse. They are the oldest examples of didactic verse, a genre that later gained enormous popularity.

The contribution of late-Vedic texts to later literature is preeminently that of the development of an expository prose style and the evolution of a sacred language, which, in order to be effective, must be completely correct. Thus, the Vedic religion evolved a science of phonetics and, later, of grammar, which was summed up in the 5th or 6th century BC by the grammarian Pāṇini in *Aṣṭādhyāyī* ("Eight Chapters"), a book that was to become basic to Sanskrit education. This language, Sanskrit, remained the language *par excellence* for later literature and was used for literary purposes until the 13th century and, epigonically, until today.

Sanskrit: epic and didactic literature (400 BC–AD 1000). After the formative period of the Vedic age, literature moved in several different directions. The close of the Vedic period was one of great cultural renewal, with the founding of the new monastic religions of Buddhism and Jainism (6th century BC) and the more slowly emerging rearticulation of Brahminism into Hinduism. Neither the earliest Buddhists nor the Jains availed themselves of Sanskrit in their preachings, apparently viewing the language as the preserve of a Brahmin elite. Sanskrit continued in derivative works of Vedic inspiration and above all in the *Mahābhārata* and the *Rāmāyaṇa*.

Mahābhārata. From references in Vedic literature it appears that side by side with the ritual texts there flourished a more secular literature carried on by bards. Originally charioteers to noblemen and thus witnesses of their feats, they chronicled the martial history of the families to which they were attached. From these beginnings, part chronicle, part panegyric, developed the epic style.

Like most Sanskrit poetry, the *Mahābhārata* consists of couplets, two successive lines with the same metre. Generally, one metre is used throughout the poem, though for stylistic effects other metres may be interspersed. The epic metre, or *śloka,* is a very fluid one that lends itself excellently to improvisation. The *Mahābhārata* is the longest poem in history, with about 100,000 couplets, more than seven times the size of Homer's *Iliad* and *Odyssey* combined. Its characters go back to around 1000 BC, but in its present form the epic could not have been composed before 400 BC. From that time until AD 400, it underwent continuous elaboration, by insertions of episodes (one of which is related in the religious poem called the *Bhagavadgītā*), accounts of separate adventures of the heroes, tales generated by their ancestors, and so on; and in the end it became a storehouse of general Hindu lore, with lengthy didactic books inserted.

The main narrative of the *Mahābhārata* recounts the growing up of two sets of cousins, both of whom aspire to a throne, the title to which is clouded. The protagonists, the Pāṇḍavas, stake their possessions in a dice game with the antagonists, the Kauravas, who are in effective control of the realm; they lose, and must live for 13 years in exile. This the five brothers do, along with the wife they hold in common. Upon their return from exile, they are refused their promised share of the kingdom, and, though parleys are held, war is inevitable. All of the Indian dynasties and tribes take sides in a war that lasts for 18 days, which only seven warriors, among them the Pāṇḍavas, survive. Noteworthy is the picture of gloom and doom that the *Mahābhārata* draws: there is little extolling of the heroic virtues of prowess and gallantry; rather, the wastefulness and bloodshed of war are pointed up, prefiguring a later concern with *ahiṃsā,* or nonviolence.

This summary does no justice to an extremely complex story with hundreds of participants, but it sketches the general outline of epic events. The main story has an unmistakable epic and heroic tone, and some of the events and encounters are completely comparable to those in epics of other peoples. But narrative and stylistic unity

are disrupted by the inserted quasi-related and unrelated secondary episodes, each of which has a style of its own, ranging from light badinage to sonorous morality tales. It was in these episodes that the *Mahābhārata* lived on and greatly influenced succeeding literature; the story of Śakuntalā, for example, which the great 5th-century classical poet Kālidāsa embroidered, the slaying of Śiśupāla, the battle of the hero Arjuna with the mountain man, the story of Nala, and so on. But the most celebrated episode surely is the *Bhagavadgītā*.

The *Bhagavadgītā*

The influence of the *Bhagavadgītā* ("Song of the Lord") has mainly been on the development of Hindu religion and philosophy. Still, it is open to doubt whether it would have exerted this influence were it not for its poetry. Like most of the *Mahābhārata*, the style is simple and direct, not given to embellishment; nevertheless, the poem often reaches the height of expressiveness, as in its evocation of the theophany of Krishna as Vishnu, in the 11th of its 18 chapters. It led to imitations such as the *Īśvaragītā*, ("Song of the Lord [Śiva]"), also in the *Mahābhārata*, in which the god Śiva (Shiva) is celebrated.

Rāmāyaṇa. While the unity of the *Mahābhārata* has been disrupted by interpolations, the unity of the second epic, the *Rāmāyaṇa*, has been remarkably preserved. It is less an epic than a romance, recounting the story of prince Rāma and his wife Sītā. The first book, a later addition, tells of the youth of the prince, who later, by the trickery of one of his father's wives, is excluded from the throne to which he is heir. He goes into voluntary exile in the forests with his wife and his brother Lakṣmaṇa. There a demon, Rāvaṇa, abducts Sītā to his island kingdom of Laṅkā. In the course of his quest for her, Rāma allies himself with a monkey nation, whose general, Hanumān, later revered as a god, discovers Sītā on Laṅkā. A monumental battle ensues. "As the sky can only be likened to the ocean and the ocean to the sky, so the battle of Rāma and Rāvaṇa can only be likened to the battle of Rāvaṇa and Rāma." After his victory, Rāma is restored to the throne, but (in what appears to be a later addition) the populace accuses Sītā of misbehaviour, probable adultery, while in Laṅkā. Rāma thus abandons her to a hermitage (the sage of the hermitage, Vālmīki, is credited with the authorship of the *Rāmāyaṇa*), where she gives birth to their twin sons. Ultimately, Rāma takes Sītā and his sons back. In the later additions, the first and probably the last books, King Rāma is accepted as an incarnation of the god Vishnu, rather than merely a perfect man and hero.

It is the main story of the romance that has made an indelible impression on Indian culture, morally as well as literarily. Rāma is the perfect, just king; Sītā, the model of an Indian wife; Lakṣmaṇa (the brother), the paragon of fraternal love; and the monkey Hanumān, the epitome of a servitor's loyalty. It was translated into and adapted in many modern Indian languages, and (like parts of the *Mahābhārata*) it found its way into Java. Vālmīki himself was hailed by later classical poets as the first true poet (*kavi*), and indeed much of his work has a poetic freshness and literary intention that is largely absent from the *Mahābhārata*. Vālmīki's great tools are metaphor and simile, as is also true of later literature. He delights in description of pastoral scenes, in lamentations and grand martial spectacles, and in the idyll of the hermitage, which depicts a serene sage leading a life of quiet meditation and living on simple forest fare in a tranquil woodland close to a sacred river. And the entire work is suffused with a confident, unwavering morality, for which the heroes of the *Mahābhārata* are still searching.

Harivaṃśa and Purāṇas. The role of the *Mahābhārata* as the storehouse of Hindu lore was supplemented by the *Harivaṃśa* ("Genealogy of Hari"—that is, the god Vishnu), which deals with the ancestry and exploits of Krishna, the Pāṇḍavas' friend and adviser in the epic but now wholly deified and identified with the great god Vishnu. Then, from perhaps the 4th century, the literature of the *Purāṇas* took over. Encyclopaedic works, often of considerable length, the *Purāṇas* deal with the mythology of time and space and of deities, with sagas of great heroic dynasties, and with legends of saints and ascetics; their interest is largely religious. Aesthetically, the most important of them is the *Bhāgavata-Purāṇa* (9th or 10th century), which celebrates the blessed lord (*bhagavat*) Vishnu in his many theophanies but is particularly evocative in its celebration of Vishnu's incarnation as Krishna and the playful story of his youth. The influence of the *Bhāgavata-Purāṇa*, particularly the 10th book, on Indian religion, art, and literature has been monumental. In the opinion of one scholar, this book constitutes the greatest poem ever written; and so it is in the popular estimation of the Hindus. It was adapted in many Indian languages and provided themes and scenes for the flourishing miniature styles of the Middle Ages.

Pāli and Prākrit literature (c. 200 BC–AD 200). No more than the Vedic literature do the literatures of early Buddhism and Jainism have a literary intention. Their texts, written in dialects other than Sanskrit, articulate the teachings of the religious founders and their successors. Because they were transmitted orally for a considerable time before they were written down in the form they would retain, they underwent the inevitable censorship of the centuries, both negative in the form of documents dropped out of use and positive in the form of newer documents added. The dates given here are only approximations of the time of the documentary fixation of the dates.

Buddhist texts. The earliest records of Buddhism are not textual but inscriptional, in the famous edicts of the Mauryan emperor Aśoka, who reigned *c.* 269–232 BC. Among these inscriptions on stone, the so-called 13th rock edict—in which Aśoka, after the massacre of the Kaliṅgas (modern Orissa), abjures war—is the most moving document of any dynastic history. The inscriptions were written in a variety of Prākrits; that is, Indo-Aryan languages closely cognate to, but considerably later than, the earliest stabilized Sanskrit.

Pāli, the language of textual literature

The vehicle of the extant textual literature is the Pāli language, which is held to be a western Indian dialect on a substratum of several central and eastern ones. It was the language in use by the Theravāda school of Buddhism; but, since that school became the dominant one among many in early Buddhism, the Pāli language is often identified with the Buddha's own speech. Most of the canonical literature is exclusively of religious interest, but interspersed in it are works of considerable literary interest.

Foremost perhaps are discourses put into the Buddha's mouth—for example, his sermon "In the Deer Park"—and no doubt deriving from fairly accurate memories. With their straightforward, lively, and incisive style, homely similes, and simple humour, they are excellent examples of the homiletics of early Buddhist preaching. Incorporated in the canon, too, are more general works of literature. The *Dhammapada* ("Verses on the Buddhist Doctrine") is a fine example of the moralistic, aphoristic strain in Indian literature, in which virtue is extolled and vice condemned. It has remained a work of considerable diffusion in all Buddhist countries, and, as in the case of the *Bhagavadgītā* in Hinduism, much of its popularity is due to its literary style. The *Suttanipāta* collection of the Buddhist canon, composed in a more formal style, contains 55 narrative and didactic poems, in the form of dialogues and ballads; they are composed in a metre akin to the Sanskrit *śloka*. Of great interest are the *Theragāthā* and the *Therīgāthā* ("Hymns of the Senior Monks" and "Hymns of the Senior Nuns"), which give at times a vivid insight into the ambience in which a conversion to Buddhism took place: a monk celebrates his newfound freedom in an idyll of the hermit's life; and a nun reminisces over the pains of deserting her home and child, yet without regrets, since she has won the freedom of Buddhism. The prosodic variety of Buddhist lyrics is great; about 30 different metres can be distinguished. Pāli poems, with their new metres (often based on a musical phrase), stylistic features, figures of speech, and choice diction, foreshadow classical *kāvya* literature in Sanskrit, whose extant specimens date from a later period.

The *Jātakas*

Of great importance is a huge volume called *Jātakas* ("Birth Stories"), recounting some 500 episodes supposedly having occurred in the Buddha's earlier lives. Only those parts in archaic verse are canonical; the prose portion was written later (*c.* 3rd century AD), probably in

Ceylon. The *Jātakas* consist of fairy tales, animal stories and fables (the future Buddha may be incarnate in an animal), ballads, and anecdotes. Though their setting is often imaginary, they provide significant material for the historian of society and culture. These mostly short tales abound in moving, delicate, often rustic touches that have made them the delight of the Buddhist world. Their themes are illustrated in bas-reliefs of Buddhist shrines (or *stūpas*) at Bhārhut and Sānchi and monumentally on the great *stūpa* of Java, the Borobuḍur.

Of considerable literary as well as historical interest is the Pāli text *Milinda-pañha* ("The Questions of Milinda"). Milinda is identical to the Greek Menander, the name of a Bactrian Indo-Greek king (*c.* 140–110 BC) who was skeptical of the verities of Buddhism and was enlightened by the teaching of an elder, Nāgasena. The extensive Buddhist erudition that the sage displays is artfully presented in the form of simile and parable, and the work has contributed importantly to the edification of audiences in the countries where Buddhism came to be established. The style, in spite of the repetitions so typical of Buddhist doctrinal texts, is lively and presents the reader with an invaluable picture of contemporary Indian life.

Jaina texts. Less interest attaches to Jaina canonical works, which were written in an adapted and stabilized literary dialect called Ardhamāgadhī (Semi-Māgadhī, Māgadhī being the dialect of the ancient kingdom of Magadha, in present day Bihār). The belletristic contribution of Jaina literature is discussed below.

Classical Sanskrit kāvya (200–1200). Prepared for by the systematization of the Sanskrit language by Pāṇini, the development of the great epics, notably the *Rāmāyaṇa,* and the refinements of prosody represented by the Pāli lyrics, there arose, in the first centuries AD, a Sanskrit literary style that governed canons of taste for a millennium and remained influential far later through modern Indian languages and their literatures. The style, called *kāvya,* is characterized by an extremely self-conscious effort on the part of the writer to compose poetry pleasing to both the ear and the mind. It evolved an elaborate poetics of figures of speech, among which the metaphor and simile, in their many manifestations, predominate; a careful use of language, governed by the stated norms of grammar; an ever-increasing tendency to use compound nouns instead of drawing on the quite plentiful possibilities of Sanskrit inflection; a sometimes ostentatious display of erudition in the arts and sciences; an adroitness in the use of varied and complicated, if appropriate, metres—all applied to traditional themes such as the epic had provided and to the rendering of emotions, most often the love between men and women.

Characteristics of the *kāvya* style

The style finds its classical expression in the so-called *mahākāvya* ("great poem"), most akin to the epyllion ("miniature epic") art form of the Alexandrian poets (a school of Greek poets, *c.* 3rd–1st centuries BC); the strophic lyric (a lyric based on a rhythmic system of two or more lines repeated as a unit); and the Sanskrit theatre. It can also be extended to narrative literature, especially the prose novel. The great masters in the *Kāvya* form (which was also exported to Java) were Aśvaghoṣa, Kālidāsa, Bāṇa, Daṇḍin, Māgha, Bhavabhūti, and Bhāravi.

The earliest surviving *kāvya* literature was written by a Buddhist, Aśvaghoṣa, said to have been a contemporary of the Kuṣāṇa (Kushān) king Kaniṣka (1st century AD). Aśvaghoṣa's work also marks a shift away from the Pāli of the Theravāda branch of Buddhism back to the more and more accepted Sanskrit of the Mahāyāna branch. Two works are extant, both in the style of *mahākāvya:* the *Buddhacarita* ("Life of the Buddha") and the *Saundarānanda* ("Of Sundarī and Nanda"). Compared with later examples, they are fairly simple in style but reveal typical propensities of writers in this genre: a great predilection for descriptions of nature scenes, for grand spectacles, amorous episodes, and aphoristic observations. The resources of the Sanskrit language are fully exploited; stylistic embellishments (*alaṅkāra*) of simile and metaphor, alliteration, assonance, and the like are employed, often quite felicitously. The original *Buddhacarita,* rediscovered in 1892, had been known from Tibetan and Chinese translations. The Sanskrit text is fragmentary, breaking off in the 14th canto (major division of the poem) with the enlightenment of the Buddha, while the other versions take the story through the Buddha's Nirvāṇa. Though intended to instruct the reader to turn away from the sensuous life and follow the Buddha's path, the work is at its best in descriptions of that very life. This is even more apparent in the *Saundarānanda,* which recounts a well-known story of how the Buddha converted his half-brother Nanda, who was deeply in love with his wife, Sundarī, and with the good life, to the monastic life of austerity. In his mastery of the intricacies of prosody and the subtleties of grammar and vocabulary, Aśvaghoṣa shows himself the complete forerunner of the Hindu *mahākāvya* authors.

The mahākāvya. In its classical form, a *mahākāvya* consists of a variable number of comparatively short cantos, each composed in a metre appropriate to its particular subject matter. The subject matter of the *mahākāvya* itself is taken from the epic, which is not, however, followed slavishly. Most *mahākāvya*s display such set pieces as descriptions of cities, oceans, mountains, the seasons, the rising of the sun and moon, games, festivals, weddings, embassies, councils, war, and triumph. It is typical of the genre that, while each strophe, or stanza, is intended to be part of a narrative sequence, it more often stands by itself, a discrete unit conveying one idea or developing one image. In this, the tendency of the Rigvedic stanza (see above *Sanskrit: formative period [1200–400 BC]*) continues in the classical literature. Although the lines of the classical stanza are long enough to convey their meaning quite explicitly, it is the pride of the poet to suggest rather than to express. Sometimes this is done by simple collocation of words: for example, in the first line of Kālidāsa's *Meghadūta* a *yakṣa* (a mischievous elf-like creature) is afflicted by a curse, "the more painful because it spelt separation from his beloved"; the next word notes that he had been negligent in his duties; taken together, the two words, though syntactically unrelated, suggest that it was his amour that made him neglect his duties. Another common suggestive device is the double meaning, or play on words. These double meanings often add a certain graceful playfulness to the poetry, reminding one that the poem was written first of all to give pleasure to the man of taste.

Traditionally there are six model *mahākāvyas,* three by Kālidāsa and one each by Bhāravi, Māgha, and Śriharṣa, to which sometimes the *Bhaṭṭikāvya* is added.

The works of Kālidāsa

Nothing is known with certainty of the life of Kālidāsa, the greatest of Sanskrit poets, but there is substantial agreement that at one time he lived in Ujjayinī (Ujjain, in the present state of Madhya Pradesh), the capital of Avanti and an important centre of Sanskrit culture in a commercially busy area. His name, which means Servitor of Kālī, indicates that he was a follower of that goddess, whom he was to celebrate as Pārvatī, the daughter of the mountain, in the *Kumārasaṃbhava.* Probably he lived during the reign of Candra Gupta II Vikramāditya (*c.* 380–*c.* 415), and there are reports that he died, by the hand of an envious courtesan, while a guest of King Kumāradāsa of Ceylon.

Compared with those of others, Kālidāsa's style might be called simple, but it is a very studied, very felicitous simplicity, hiding the actual complexity of his constructions. In two of his *mahākāvyas,* Kālidāsa draws on epic lore. The first, and probably earlier one, is the *Kumārasaṃbhava* ("Birth of the War God"), which describes the courting of the ascetic Śiva, who is meditating in the mountains, by Pārvatī, the daughter of the Himalayas; the destruction of the god of love (after his arrow has struck Śiva) by the fire from Śiva's third eye; and the wedding and lovemaking of Śiva and Pārvatī, which results in the conception of the war god. The original is in eight cantos, but a sequel was added by an imitator. The second *mahākāvya,* the *Raghuvaṃśa* ("Dynasty of Raghu"), deals with themes from the *Rāmāyaṇa:* it describes the vicissitudes of the Solar dynasty of the ancient Indian barons, culminating in the *Rāmāyaṇa* story of Rāma and Sītā. The *Raghuvaṃśa* is famous for its beautiful descriptions and incidental narratives, which give the poem a somewhat episodic character; among them are a description of

the six seasons (spring, summer, rainy, autumn, winter, and dewy) and the story of a young hermit who went to the river to fill a water jar for his parents and was killed by a stray arrow.

Unique in Sanskrit love poetry is Kālidāsa's *Meghadūta,* in which the poet tries to go beyond the strophic unity of the short lyric (see below), which normally characterizes love poems, by stringing the stanzas into a narrative. This innovation did not take hold, though the poem inspired imitations along precisely the same story line. The *Meghadūta* is the lament of an exiled *yakṣa* who is pining for his beloved on a lonely mountain peak. When, at the beginning of the monsoon, a cloud perches on the peak, he asks it to deliver a message to his love in the Himalayan city of Alakā. Most of the poem, composed in an extremely graceful metre, consists of a description of the landmarks, cities, and the like on the cloud's route to Alakā. It must be considered among the finest poems, if not the finest poem, written in Sanskrit. Kālidāsa also wrote for the theatre (see below) and was no doubt the most versatile author of Sanskrit literature; his works became well-nigh canonical models.

Bhāravi (6th century) probably hailed from the south during the reign of the Pallava dynasty. He took up a *Mahābhārata* theme in his *Kirātārjunīya* ("Arjuna and the Mountain Man"), recounting the Pāṇḍava prince Arjuna's encounter and ensuing combat with a wild mountaineer who in the end proves to be the god Śiva. Bhāravi's language and style are more difficult than Kālidāsa's, but the poem is highly regarded in Indian literary tradition.

Māgha, who wrote in the 8th century, was a conscious rival of Bhāravi, whom he attempted to surpass in every respect. His *Śiśupālavadha* ("The Slaying of King Śiśupāla") is based on an episode of the *Mahābhārata* in which the rival King Śiśupāla insults the hero-god Krishna, who beheads him in the ensuing duel. Māgha is a master of technique in the strict Sanskrit sense of luscious descriptions; intricate syntax; compounds that, depending on how they are split, deliver quite different meanings; and the full register of stylistic embellishments.

To some critics, the preoccupation with technique, the triumph of form over substance, appears to have spelled the doom of the *mahākāvya.* A curious but entirely Sanskritic phenomenon, for example, is the *Bhaṭṭikāvya,* a poem by Bhaṭṭi (probably 6th or 7th century). It again deals with the story of Rāma and Sītā, but at the same time it illustrates in stanza after stanza, in exactly the proper sequence, the principal rules of Sanskrit grammar and poetics. Less artificial is the *Naiṣadhacarita* ("The Life of Nala, King of Niṣadha"), written by the 12th-century poet Śrīharṣa and based on the story of Nala and Damayantī in the *Mahābhārata.* An example of another kind of excess indulged in by *mahākāvya* writers is the *Rāmacarita* ("Deeds of Rāma"), by the 12th-century poet Sandhyākāra, which celebrates simultaneously the hero-god Rāma and the poet's own king, Rāmapāla of Bengal. Many other works were written in this style, and, even today, one may encounter a *mahākāvya* treatment of a great man such as Mahatma Gandhi or Jawaharlal Nehru.

Jayadeva's *Gītagovinda*

Difficult to classify is the work of the 12th-century Bengal poet Jayadeva, who wrote the *Gītagovinda* ("Cowherd Song"). The basic structure of this long poem, in which the poet recounts the youthful loves of the cowherd hero and god Krishna, largely based on the story of the *Bhāgavata-Purāṇa,* is that of the *mahākāvya.* Generously interspersed between cantos, however, are erotic-religious lyrics of extremely musical assonances, which were, and still are, sung. Jayadeva's work, rather lacking in the grammatical rigidity of the other *mahākāvya* writers, has been extremely popular and affords a fine example of the devotional lyric (see below).

The short lyric. It is in the short, one-stanza lyric that Sanskrit poetry is revealed most intimately in its real aims. As noted, almost all of high Sanskrit poetry is strophic in fact; in the lyric it is so in intention. It is eminently a genre of the poetic moment, making an aesthetic observation and placing it within the Sanskritic universe of discourse. It may be an observation of anything: a fish glintingly jumping from a pond, aboriginal tribesmen engaged in a bloody rite, love in all its manifestations, a glimpse of God perceived or remembered. But in the monumental lyric collections that have been preserved, and in the many stray verses still circulating among educated Hindus in India as so-called *subhāṣita*s ("well-turned" couplets), the more common topics are praise of the god of one's devotion and the vagaries of love.

In the short lyric it is hard to make a distinction that depends on the language in which it is composed; for, although the language may be different, the subject matter and forms are the same. Many love lyrics, especially when they describe feelings experienced by women, are composed not in Sanskrit but, instead, in one of the Prākrits, or Middle Indo-Aryan languages, among which the dialect called Māhārāṣṭrī is particularly popular. The collection of 700 poems in this language, compiled by Hāla under the name of *Sattasaī* ("The Seven Hundred"), tends to be simpler in imagery and in the emotion portrayed than their Sanskrit counterparts, but essential differences are difficult to pinpoint.

The devotional lyric, a short verse expressing the author's devotion to a god, is linked with both the hymnal poetry of the Rigveda—though far less determined by a desire for compelling magic—and the temple worship of Hinduism. Though by no means always, there is often a particularism about them: the deity is invoked as it appears in a specific iconic stance or in a local temple or in a manifestation especially pleasing to the poet. The number of such verses is countless; every major religious and philosophic leader is held to have added to their stock. Some are especially famous: the *Sūryāṣṭaka* ("Eight Strophes for the Sun"), by Mayūra; the collections attributed to the philosopher Śaṅkara, the *Saundaryalaharī* ("The Wavy River of the Beautiful Sky"); and the *Kṛṣṇakarṇāmṛta* ("The Elixir of Hearing of Krishna"), by Bilvamaṅgala, among others. These *stotra* ("lyrics of praise") quite often were set to music, and people continue to sing them today—without necessarily comprehending the full intention of the Sanskrit, much as hymns in Latin were traditionally sung by Roman Catholic believers.

The love lyric

The entire erotic experience, from budding love to the aftermath of consummation, is represented brilliantly in lyric poetry. But among the many themes inspired by love, poets have been most attracted to the lament of separated lovers. It is mostly the sufferings of the woman that are portrayed, but the grief of the man is also depicted—in Kālidāsa's *Meghadūta,* for example. The love lyrics consist of single verses, many of which seek to suggest the mood of *śṛṅgāra* (physical love). While often extremely erotic, they are very rarely obscene. Sanskrit norm banned all coarse expressions for sexual play; and, although much probably escapes the modern reader, blunt allusions to genital organs are rare and, where allusions occur, extremely veiled. Bodily parts with less overt sexual connotations, such as breasts and buttocks, are frankly mentioned and described—in fact, celebrated. In allusions to sexual intercourse the terminology of the *Kāmasūtra* of Vātsyāyana is frequently invoked, as though this ancient textbook of Indian erudition was a protection against possible opprobrium—not unlike Latin terms resorted to in the West for actions that most know by shorter, more colloquial names.

The erotic and the devotional lyric merge freely, and at times it is impossible to make out whether the free sexual imagery employed is to be taken literally or as an allegory of the human soul courting the love of its god. The task—not a very pressing one—is made more difficult by the fact that some *bhakti* (devotion) religions have developed the poetics of love poetry into a kind of theology, a phenomenon quite characteristic of Bengal Krishnaism (see below *Indo-Aryan literatures: 12th–18th century*).

Authors of *subhāṣita*s often collected them themselves, the favourite form being that of the *śataka* ("century" of verses), in which 100 short lyrics on a common theme were strung together. Mention has been made of Hāla's *Sattasaī* ("The Seven Hundred," consisting of lyrics in the Māhārāṣṭrī dialect). Four well-known Sanskrit collections, of the 7th century, are the famous "century" of Amaru, king of Kashmir, and the three "centuries" by the poet

Bhartṛhari; one of the latter's collections is devoted to love, another to worldly wisdom—a very popular theme in epigrammatic verse—and the third to dispassion. Of the same type but in a different vein is *Caurapañcāśikā* ("Fifty Poems on Secret Love"), in which the 12th-century poet Bilhaṇa fondly recalls the pleasure of his clandestine amours with a local princess.

The theatre. Of all the literary arts, the Indians esteemed the play most highly, and it is in this form that most of the other arts were wedded together. Its origins are obscure, but there is reason to assume that the play developed out of recitations of well-known epic stories by professional reciters. It is an extremely rich genre with a number of outstanding playwrights.

The style is extremely varied. Although it might be called a Sanskrit play, Sanskrit is by no means the only language used, for the less educated characters, including all women, speak Prākrits of different degrees of niceness. The action is carried by prose, but at the least provocation—indeed, at any of the poetic moments characteristic of the strophic lyric—the author reverts to verse, sometimes in mid-sentence. Two principal types of play are distinguished: the *nāṭaka,* which is based on epic material, and the *prakaraṇa,* which is of the author's invention, though often borrowed from narrative literature.

Principal types of drama

Characteristic of the Sanskrit theatre are elements of sacrality. The play begins and ends with a benediction, many of which consist of subject matter taken from sacred texts. It is also expressed in numerous taboos: the play must have a happy outcome in which harmony, interrupted by the drama of the play, is restored; improper scenes, such as eating, dressing and undressing, and sexual intercourse, are not to be portrayed; no violence among the higher characters is permitted; war, which often occurs, should simply be reported on, often by lower characters, not in any way staged.

Fragments of Buddhist plays prior to the flowering of Hindu theatre have survived, but no complete plays earlier than 13 ascribed to the playwright Bhāsa. There is considerable controversy over the authenticity of the Bhāsa plays, but at least some of them must be authentic, perhaps dating back to the 3rd century. The plays are based on the epic and on the *Bṛhat-kathā* narrative cycle (see below); among the latter, the *Svapnavāsavadattā* ("The Dream of Vāsavadattā") is the most famous. Of considerable interest also is the *Daridra-Cārudatta* ("The Poverty of Cārudatta"), which became the basis for the play *Mṛcchakaṭika* ("Little Clay Cart") of Śūdraka (see below).

It must be assumed that there was an efflorescence of poetry and theatre in the city of Ujjayinī, one of the capitals of the Gupta Empire, in the 5th century, for a number of authors can be placed there during this reign; among these were Viśākhadatta, Śūdraka, Śyāmilaka, the writer of one of the best farces, and Kālidāsa, who at the beginning of the development of the genre produced some of the greatest plays in the tradition.

Kālidāsa's plays

Three plays by Kālidāsa remain, one of which is the *Mālavikāgnimitra* ("Agnimitra and Mālavikā"), a harem play of amorous intrigue at a royal court. The other two are based on old themes. *Vikramorvaśī* ("Urvaśī Won by Valour") is based on a story as old as the Rigveda, that of the nymph Urvaśī, who is loved by King Purūravas, whom she marries on the condition that she shall never see him nude. The accident happens, and the nymph returns to heaven, leaving her husband crazed with longing, until a final reunion. But the Indian tradition holds the *Abhijñānaśakuntalā* ("Śakuntalā and the Token of Recognition") to be the greatest of all Sanskrit plays. It recounts a *Mahābhārata* story—rather freely to be sure—of a hermit girl secretly married to a visiting king, who leaves with her a keepsake that will serve her as a token of recognition. She gives birth to a son, Bharata, and goes to the King's court; on the way she loses the ring in a river, where a fish swallows it. The King fails to recognize her and rejects her, and her mother, a nymph, carries her to heaven. When the ring is recovered by a fisherman and the King's memory is restored, he searches for Śakuntalā but does not find her. In the end he meets a boy who proves to be his son and is restored to him.

Kālidāsa's great forte is the portrayal of emotions—ordinary enough in themselves (budding love, love consummated, rejection, despair, a father's love for his son)—but Kālidāsa applies to them a mastery of expression and image that makes the play a work of perennial beauty.

Next to nothing is known of Śūdraka except that he must have hailed from Ujjayinī. His is the most charming of all *prakaraṇa* plays (those that are not based on epic material): the *Mṛcchakaṭikā* ("Little Clay Cart"), the story of an impoverished merchant and a courtesan who love each other but are thwarted by a powerful rival who tries to kill the woman and place the blame on the hero, Cārudatta. The play offers a fascinating view of the different layers of urban society. Viśākhadatta, the author of a rare semi-historical play called *Mudrārākṣasa* ("Minister Rākṣasa and his Signet Ring"), apparently was a courtier at the Gupta court. His play is a dramatization of the Machiavellian political principles expounded in the book *Arthaśāstra,* by Kauṭilya, who appears as the hero of the play.

To the 7th-century king Harṣa of Kanauj are attributed three charming plays: *Ratnāvalī* and *Priyadarśikā,* both of which are of the harem type; and *Nāgānanda* ("The Joy of the Serpents"), inspired by Buddhism and illustrating the generosity of the snake deity Jīmūtavāhana.

The plays of Bhavabhūti

Ranked by Indian tradition close to Kālidāsa himself, Bhavabhūti (early 8th century) was the author of three plays, two of which are based on the *Rāmāyaṇa* story. The *Mahāvīracarita* ("The Exploits of the Great Hero") treats of Rāma's battle with Rāvaṇa and the *Uttararāmacarita* ("The Later Deeds of Rāma") treats of the life of Rāma after he has abandoned Sītā. Bhavabhūti lacks the elegance and grace of Kālidāsa but is more pensive—even brooding—than his predecessor. His style is also very forceful. His *prakaraṇa Mālatī-Mādhava* ("Mālatī and Mādhava") is a complex love intrigue intermingled with sorcery and Tantric practices, including a human sacrifice and much violence.

This list by no means concludes that of the playwrights in the Sanskrit tradition. The writing of plays, mostly derivative from the great models, has continued until the present day.

Apart from the more seriously intended plays described above, the Sanskrit theatre also has a rich repertory of farces, which are usually in one act. Most interesting of these are the *bhāṇas,* which may be monologues in which an actor addresses imaginary persons and is answered by them, as he paints a picture of town life full of personal and social satire. Among the best in this little-studied genre is Śyāmilaka's 5th-century *Pādataḍitaka* ("The Courtesan's Kick").

Narrative literature. Sanskrit narrative literature is extremely rich, so rich in fact that at one time it was believed that all folktales originally came from India. Many indeed have, and they have found a place in *The Arabian Nights' Entertainment,* Boccaccio's *Decameron,* and other such works down to the fairy tales of Hans Christian Andersen and the fables of Jean de La Fontaine. Certain collections of animal tales, some of which go back to the Buddhist *Jātaka* stories, had incredible histories. The most famous is the *Pañca-tantra* ("The Five Chapters"), which, within a framework of a lesson in the art of politics addressed to young princes, presents a number of animal characters who in their actions both admonish and exhort the reader to a life certain to lead to worldly success. A shorter version, partly drawn from the *Pañca-tantra,* is the *Hitopadeśa* ("Good Advice"). The *Pañca-tantra* found its way to the West through translations into Persian, Arabic, Syrian, Hebrew, and Latin, until most of the medieval literatures possessed their own versions of it. No less extensive were its migrations to Southeast and East Asia.

The *Pañca-tantra*

The principal work of the novelistic and picaresque tale is the *Bṛhat-kathā* ("Great Story") of Guṇāḍhya, written in Prākrit and now lost, save for Sanskrit retellings. The most important among these Sanskrit versions is the *Kathā-saritsāgara* ("Ocean of Rivers of Stories") of Somadeva (11th century), which includes so many subsidiary tales that the main story line is frequently lost. Perhaps more faithful to the original—in any case far less distracting—is the *Bṛhatkathāślokasaṃgraha* ("Summary in Verse of the

Great Story"), by Budhasvāmin (probably 7th century), one of the most charming of Sanskrit texts. Other collections of tales include the *Vetāla-pañcaviṃśatikā* ("Twenty-five Tales of a Ghost"), *Śūkasaptati* ("The Seventy Stories of a Parrot"), and the *Siṃhāsana-dvātrim-sātikā* ("Thirty-two Stories of a Royal Throne").

Related to the *Bṛhat-kathā* cycle, though the exact relationship is unclear, is the Jain Prākrit text of the *Vāsudevahiṇḍī*, "The Roamings of Vāsudeva" (before 6th century), describing the acquisition of numerous wives by Krishna Vāsudeva.

Though the tales are often artless, sometimes they are elaborately embroidered in the Sanskrit *kāvya* style. A fine example is the *Daśakumāracarita* ("Tales of Ten Princes"), by Daṇḍin (6th/7th century), in which, within the framework of a boxing story, the picaresque adventures of 10 disinherited princes are described in prose. While tending overly to description, the work remains eminently readable for the modern reader, a quality that cannot be attributed to the prose novels of the 7th-century writer Bāṇa: the *Harṣacarita*, "The Life of Harṣa" (king of Kanauj and the author of three plays, discussed above), which is important for its information on culture and society; and the *Kādambarī* (the name of the heroine), which describes the affairs of two sets of lovers through a series of incarnations, in which they are constantly harassed by a cruel fate. (J.A.B.v.B.)

DRAVIDIAN LITERATURE: 1ST–19TH CENTURY

Of the four literary Dravidian languages, Tamil has been recorded earliest, followed by Kannada, Telugu, and Malayalam. Tamil literature has a classical tradition of its own, while the literatures of the other languages have been influenced by Sanskrit models.

Early Tamil literature (1st–10th century). *Caṅkam literature.* Early classical Tamil literature is represented by eight anthologies of lyrics, 10 long poems, and a grammar called the *Tolkāppiyam* ("Old Composition"). According to a fanciful Tamil tradition, this literature was produced by poets of three "academies," or *caṅkams*, that in the hoary past were centred in the southern Indian city of Madurai and supposedly lasted 4,400, 3,700, and 1,850 years, respectively. The *Tolkāppiyam* was ascribed to the second *caṅkam*, the eight anthologies and 10 long poems to the third; according to tradition, nothing is extant from the first *caṅkam*. The early literature, itself known as *caṅkam*, comprises 2,381 poems, ranging from four to nearly 800 lines each and assigned to 473 poets who are known by name or epithet; about 100 poems are anonymous. Though the literature does not go back as far as native tradition would have it, it is generally ascribed to the first three centuries of the Christian Era and represents the oldest non-Sanskrit literature to be found on the South Asian subcontinent.

The eight anthologies and their contents, excluding opening invocations that were added later, are as follows: *akam* anthologies consisting of (1) *Kuṟuntokai*, 400 love poems, (2) *Naṟṟiṇai*, 400 love poems, (3) *Akanāṉūṟu*, 400 love poems, (4) *Aiṅkuṟunūṟu*, 500 love poems, each 100 (assigned to a different poet) dealing with one of five phases of love, (5) *Kalittokai*, 150 love poems in a metre called *kali;* and *puṟam* anthologies consisting of (6) *Puṟanāṉūṟu*, 400 poems, (7) *Patiṟṟuppattu* ("The Ten Tens"), 100 poems on kings (the first and last decades are missing), and (8) *Paripāṭal*, a collection of 70 religious poems. *Paripāṭal* and *Kalittokai* appear to be the latest of the anthologies; *Kuṟuntokai* and *Puṟanāṉūṟu* probably contain the earliest compositions. The remarkable work of grammar and rhetoric, *Tolkāppiyam*, is the crucial text for an understanding of early Tamil language and literature. Divided into three sections (each consisting of *cūṭtirams*, or aphorisms)—sounds, words, and meaning—the *Tolkāppiyam* details, in the third, the canons of *caṅkam* poetic traditions.

Characteristics of *caṅkam* poetry

In the *Tolkāppiyam* and the anthologies, poems are classified by theme into *akam* ("interior") and *puṟam* ("exterior"), the former highly structured love poems, the latter heroic poems on war, death, personal virtues, the ferocity and glory of kings, and the poverty of poets. Both the *akam* and the *puṟam* had well-defined *tiṇai*s (genres) that paralleled one another: *e.g.*, the *kuṟiñci* genre, in love poetry, which dealt with the lovers' clandestine union on a hillside by night; and the *veṭci* genre, in heroic poetry, which dealt with the first onset of war, by nocturnal cattle stealing. Both *kuṟiñci* and *veṭci* are names of flowers that grow on the hillside, here symbolic of the poetic genre, the mood, and the theme. By such pairings across *akam* and *puṟam*, love and war become part of the same universe and metaphors for one another; the same poets—for example, Paraṇar and Kapilar—wrote great poems in both genres. The basic technique depended on a taxonomy of Tamil nature and culture, of culturally defined time, space, nature, and human nature. For example, matched in metaphor with five phases of *akam* love (union; infidelity; anxious waiting; patient waiting; and the lover or lovers eloping or journeying for wealth, knowledge, and so on) are six seasons, six parts (dawn, forenoon, noon, afternoon, evening, and night) of the day, and five landscapes (hill, seashore, forest, pasture, and wasteland, named after characteristic flowers—*kuṟiñci, neytal, mullai,* and *marutam*—and the evergreen tree, *pālai*) and their contents (including gods, foods, birds, beasts, drums, occupations, lutes, musical styles, flowers, and kinds of running or standing water). Each landscape becomes a repertoire of images—anything in it, bird or drum, tribal name or dance, may evoke a specific feeling. A favourite poetic device is *uḷḷuṟai* (*i.e.*, metonymy, a figure of speech consisting of the description of one thing used to evoke that of another with which it is associated). Thus, the natural scene implicitly evokes the human scene; for example, bees making honey out of *kuṟiñci* flowers evokes the lovers' union. Not only is the poet's language Tamil, but the landscapes, the personae, and the appropriate moods and situations formulate the realities of the Tamil world into a code of symbols. For some five or six generations, the *caṅkam* poets spoke this common language of symbols, creating a body of lyrical poetry probably unequalled in passion, maturity, and delicacy by anything in any Indian literature.

Eighteen Ethical Works. The *Patiṟeṇ-kīṛkkaṇakku* ("Eighteen Ethical Works"), usually dated as post-*caṅkam* (4th–7th centuries), are all affected by Jainism and Buddhism. Of these the *Tirukkuṟaḷ* ("Sacred Couplets"), ascribed to Tiruvaḷḷuvar, is the most celebrated. Its 1,330 hemistichs (half lines of verse) are probably the final distillation of different periods. There are many parallels in the work to the Sanskrit *Kāma-sūtra*, the treatise on erotic love, to *Manu-smṛti*, an ancient treatise on special obligation and religious law, and to *Artha-śāstra*, Kauṭilya's treatise on politics. The *Kuṟaḷ* has three sections: *aṟam*, or virtue (Sanskrit *dharma*); *poruḷ*, government and society (Sanskrit *artha*); and *kāmam*, love (Sanskrit *kāma*). There is no special treatment of *mokṣa*, or salvation, though *aṟam* seems to include it. In the *aṟam* (virtue) section, the *Kuṟaḷ* sums up a world-affirming wisdom, the wisdom of human sympathy, expanding from wife, children, and friends to clan, village, and country. In the *poruḷ* (government and society) section, the aphorisms project a vision of an ideal state, based on educated human nature, and relate the good citizen to the good man. Prostitution, disease, drink, and gambling are listed, with foreign enemies, as dangers to the state. In the *kāmam* (love) section, the *Kuṟaḷ* follows the *caṅkam's* love—eros, or sexual love—yet anticipates agape, the perfecting of love through many lives, which appears in religious poetry of the next age.

Epics. The age of the Pallavas (300?–900), a warrior dynasty of Hindu kings, is known for its epics, beginning with *Cilappatikāram* ("The Jewelled Anklet") and *Maṇimēkalai* ("The Girdle of Gems") and including an incomplete narrative, *Peruṅkatai* ("The Great Story"), the *Cīvakacintāmaṇi* ("The Amulet of Cīvakaṉ") by Tiruttakkatēvar, and *Cūḷāmaṇi* ("The Crest Jewel") by Tōlāmoḻittēvar. The last three works depict Jaina kings and their ideals of the good life, nonviolence, and the attainment of salvation through self-sacrifice. They are also characterized by excellent descriptions of city and country and by a mixture of supernatural and natural elements. In their episodic methods of narration and set descriptions of erotic, heroic, and religious themes, these Jaina

epics became both models and sources for later epic works.

The *Cilappatikāram*

The *Cilappatikāram,* by Iḷaṅkō Aṭikaḷ, is in three books, set in the capitals of the three Tamil kingdoms: Pukār (the Cōḷa capital), Maturai (*i.e.,* Madurai, the Pāṇṭiya [Pāṇḍya] capital), and Vañci (the Cēra capital). The story is not about kings but about Kōvalaṉ, a young Pukār merchant, telling of his marriage to the virtuous Kaṇṇaki, his love for the courtesan Mātavi, and his consequent ruin and exile in Maturai, where he dies, unjustly executed when he tries to sell his wife's anklet to a wicked goldsmith who had stolen the Queen's similar anklet and charged Kōvalaṉ with the theft. Kaṇṇaki, the widow, comes running to the city and shows the King her other anklet, breaks it to prove it is not the Queen's—Kaṇṇaki's contains rubies, and the Queen's contains pearls—and thus proves Kōvalaṉ's innocence. Kaṇṇaki tears off one breast and throws it at the kingdom of Maturai, which goes up in flames. Such is the power of a faithful wife. The third book deals with the Cēra king's victorious expedition to the north to bring Himalayan stone for an image of Kaṇṇaki, now become a goddess of chastity (*pattiṉi*).

The *Cilappatikāram* is a fine synthesis of mood poetry in the ancient Tamil *caṅkam* tradition and the rhetoric of Sanskrit poetry—even the title is a blend of Tamil and Sanskrit—including in the epic frame *akam* lyrics, the dialogues of *Kalittokai* (poems of unrequited or mismatched love), chorus folk song, descriptions of city and village, lovingly technical accounts of dance and music, and strikingly dramatic scenes of love and tragic death. One of the great achievements of Tamil genius, the *Cilappatikāram* is a detailed poetic witness to Tamil culture, its varied religions, town plans and city types, the commingling of Greek, Arab, and Tamil peoples, and the arts of dance and music.

Maṇimēkalai (the heroine's name, "Girdle of Gems"), the second, "twin," epic (the last part of which is missing), by Cātaṉār, continues the story of the *Cilappatikāram;* the heroine is Mātavi's daughter, Maṇimēkalai, a dancer and courtesan like her mother. Maṇimēkalai is torn between her passion for a princely lover and her spiritual yearnings, the first encouraged by her grandmother, the second by her mother. She flees the attentions of the prince, and, while he pursues her, she attains magical powers: she changes forms; survives prison, lecherous villains, and other dangers; converts the Queen; and finally goes to Pukār, which is being destroyed by oceanic erosion, worships Kaṇṇaki, and arrives in Vañcī to work in famine relief and to perform "penance." Unlike the *Cilappatikāram,* the *Maṇimēkalai* is partisan to Buddhism. It is known for its poetry and its lively discussions of religion and philosophy.

Bhakti poetry. From the 6th century onward, a movement with religious origins made itself heard in literature. The movement was that of *bhakti,* or intense personal devotion to the two principal gods of Hinduism, Śiva and Vishnu. The earliest *bhakti* poets were the followers of Śiva, the Nāyaṉārs (Śiva Devotees), whose first representative was the poetess Kāraikkāl Ammaiyār, who called herself a *pēy,* or ghostly minion of Śiva, and sang ecstatically of his dances. Tirumūlar was a mystic and reformer in the so-called Siddhānta (Perfected Man) school of Śaivism, which rejected caste and asceticism, and believed that the body is the true temple of Śiva. There were 12 early Nāyaṉār saints. Similar poets, in the tradition of devotion to the god Vishnu, also belonged to this early period. Called Āḻvārs (Immersed Ones), they had as their first representatives Poykai, Pūtaṉ, and Pēyār, who composed "centuries" (groups of 100) of linked verses (*antāti*), in which the final line of a verse is the beginning line of the next and the final line of the last verse is the beginning of the first, so that a "garland" is formed. To these Āḻvārs, God is the light of lights, lit in the heart.

The Nāyaṉārs

The most important Nāyaṉārs were Appar and Campantar, in the 7th century, and Cuntarar, in the 8th. Appar, a self-mortifying Jain ascetic before he became a Śaiva saint, sings of his conversion to a religion of love, surprised by the Lord stealing into his heart. After him, the term *tēvāram* ("private worship") came to mean "hymn." Campantar, too, wrote these personal, "bone-melting" songs for the common man. Cuntarar, however, who sees a vision of 63 Tamil saints—rich, poor, male, female, of every caste and trade, unified even with bird and beast in the love of God—epitomizes *bhakti.* To him and other Bhaktas, every act is worship, every word God's name. Unlike the ascetics, they return man to the world of men, bringing hope, joy, and beauty into religion and making worship an act of music. Their songs have become part of temple ritual. Further, in *bhakti,* erotic love (as seen in *akam*) in all its phases became a metaphor for man's love for God, the lover.

In the 9th century, Māṇikkavācakar, in his great, moving collection of hymns in *Tiruvācakam,* sees Śiva as lover, lord, master, and guru; the poet sings richly and intimately of all sensory joys merging in God. Minister and scholar, he had a child's love for God.

Āṇṭāḷ (8th century), a Vaiṣṇava poetess, is literally lovesick for Krishna. Periyāḻvār, her father, sings of Krishna in the aspect of a divine child, originating a new genre of celebrant poetry. Kulacēkarar, a Cēra prince, sings of both Rāma and Krishna, identifying himself with several roles in the holy legends: a *gopī* in love with Krishna or his mother, Devakī, who misses nursing him, or the exiled Rāma's father, Daśaratha. Tiruppāṇāḻvār, an untouchable poet (*pāṇaṉ*), sang 10 songs about the god in Śrīraṅgam, his eyes, mouth, chest, navel, his clothes, and feet. To these Bhaktas, God is not only love but beauty. His creation is his jewel; in separation he longs for union, as man longs for him. Tirumaṅkaiyāḻvār, religious philosopher, probably guru (personal religious teacher and spiritual guide in Hinduism) to the Pallava kings, and poet of more than 1,000 verses, was apparently responsible for the building of many Vaiṣṇava temples. The last of the Āḻvārs, Nammāḻvār (Our Āḻvār), writing in the 9th century, expresses poignantly both the pain and ecstasy of being in love with God, revivifying mythology into revelation.

Period of the Tamil Cōḷa Empire (10th–13th century). The next period, the time of the Tamil Cōḷa Empire (10th–13th centuries), saw an awakening of neighbouring literatures: Kannada, Telugu, and Malayalam. The first extant Kannada work is the 9th-century *Kavirājamārga* ("The Royal Road of Poets"), a work of rhetoric rather indebted to Sanskrit rhetoricians, containing the first descriptions of the Kannada country, people, and dialects, with references to earlier works. From the 10th century on, *campū* narratives (part prose, part verse) became popular both in Kannada and in Telugu, as did renderings of the Sanskrit epics *Rāmāyaṇa* and *Mahābhārata* and Jaina legends and biography.

In Kannada, this period was dominated by the "three gems" of Jaina literature, Pampa, Ponna, and Ranna, as well as by Nāgavarma I, a 10th-century Kannada grammarian. Pampa was the *ādikavi* ("first of poets"), having attained that stature with two great epics: *Vikramārjuna Vijaya* and *Ādipurāṇa.* The former is a rendering of the *Mahābhārata,* with the hero, Arjuna, identified with the poet's royal patron, Arikēsari. This felicitous epic is known for its succinct, powerful characterizations, its rich descriptions of Kannada country and court, its moving sentiments, and its harmonious blend of Sanskrit and Kannada. While the *Vikramārjuna* is a secular work, Pampa's *Ādipurāṇa* tells the story of the Jaina hero-saint Purudēva, his previous lives, his life from birth to marriage to holy death, as well as the lives of his sons, Bharata and Bāhubali.

Telugu had its *ādikavi* ("first of poets"), in the Brahmin Nannaya Bhaṭṭa (1100–60), who, in *campū* style, wrote three books of a version of the *Mahābhārata,* later finished by Tikkana (13th century) and by Errāpraggaḍa. Like other regional versions of the *Mahābhārata,* the Telugu version is not a literal translation but an interpretation, with many local elements and differences of emphasis; for example, Nannaya emphasizes the importance of Vedic religion. Such works have made the Sanskrit epics and *Purāṇas* part of a live and growing tradition, both oral and literary, in the regional language.

The work of Kampaṉ

This period also saw the eminence of Kampaṉ's Tamil version of the *Rāmāyaṇa* (12th century). In him there is

a climactic blend of earlier *caṅkam* poetry, Tamil epics, the Āḻvārs' fervour of personal *bhakti* (devotion) toward Rāma, folk motifs, and Sanskrit stories, metres, and poetic devices. Instead of a just king and a perfect man, Rāma is an incarnation of Vishnu and an intense object of devotion, dwarfing the Vedic gods; Kampaṉ called his work *Irāmāvatāram* ("Rāma's Incarnation"); yet the emphasis is not on Vishnu but on *dharma* ("the law"), localized and Tamilized. More like Sanskrit than *caṅkam* poets, Kampaṉ revels in elaborate metaphor, hyperbole, and fanciful descriptions of virtue and nature. The work is long, consisting of about 40,000 lines; the *Yuttakāṇṭam* ("War Canto") alone, with 14 battles, equals the *Iliad* in length. The poem is also justly known for its variety of style, its exploitation of the resources of Tamil and Sanskrit both in form and content, its humour, and its handling of the narrative, dramatic, and lyric modes.

Kampaṉ's popularity extended not only into all of Tamil country but apparently into the north, influencing some episodes of Tulsī's Hindi version of the *Rāmāyaṇa*, and into northern Kerala, where 32 plays based on Kampaṉ are enacted ritually with marionettes in Śiva temples.

Pre-15th-century Tamil influence on early Malayalam, the language of Kerala, was strong and led to the literature of *pāṭṭu* ("song"), in which only Dravidian, or Tamil, phonemes may occur and Tamil-like second-syllable rhymes are kept. The best known *pāṭṭu* is *Rāmacaritam* (*c.* 12th–13th century; "Deeds of Rāma"), probably the earliest Malayalam work written in a mixture of Tamil and Malayalam. Other *pāṭṭus* in Tamilized Malayalam, written by a family of poets (14th–15th centuries) from Niraṇam in central Travancore, appear in *Kaṇṇassan Pāṭṭukaḷ*, in which Tamil conventions of metre and phonology are loosened and more Sanskrit is allowed. Similar in style is a version of the *Rāmāyaṇa* by Rāma Paṇikkar, an abridged *Bhagavadgītā* by his uncle Mādhava Paṇikkar, and a condensed *Mahābhārata* and the 10th book of the *Bhāgavata-Purāṇa* by another uncle, Śaṅkara Paṇikkar.

As strong as Tamil influence was, the predominant influence on Malayalam was Sanskrit, in language as well as literary form. The influence on language led early to a mixture of Sanskrit and Malayalam in a literary dialect called *maṇipravāḷa* (meaning "necklace of diamonds and coral"). The author of the *Līlātilakam*, a 14th-century treatise on grammar and poetics, describes both the Tamilizing and Sanskritizing trends and genres and insists on harmonious blendings. Many kinds of poems were composed in *maṇipravāḷa* styles: *kūḍiyāṭṭams* (dramatic presentations using Sanskrit *ślokas*, or epic metres, for hero and heroine, *maṇipravāḷa* for the clown, and Malayalam for explanations intended for the laity); didactic works such as the 11th-century *Vaiśikatantram* ("Advice to a Courtesan by Her Mother"); 13th- and 14th-century *campūs* (narratives combining prose and verse) on dancers, such as *Unniyati Caritam* by Dāmōdara Cākkiyār; and several short poems in praise of women and kings. *Maṇipravāḷa* poems like these are essentially artifical expressions of courtly high-caste poets, preoccupied with eroticism and harlots. The *Candrōtsavam* (*c.* 1500; "Moon Festival") is a satire on the voluptuary *maṇipravāḷa* tradition, jostling together all the famed courtesans of the period.

Coexisting with the Tamilized and Sanskritized Malayalam poems produced by scholars was a live *pacca* ("pure, fresh") Malayalam tradition represented mostly by folk songs and ballads—for example, *Vaḍukkan Pāṭṭukaḷ* (hero ballads of the northern Malabar Coast); songs sung during weddings, deaths, or festivals; and work songs. All three styles—the indigenous folk style, the Tamil, and the Sanskrit—began to converge and influence each other by the 15th century in works such as *Kṛṣṇa Pāṭṭu* ("Song of Krishna") and *Gāthā* ("Song"). Such grafting reached its full flowering in the 16th-century poet Eḻuttaccan (Father [or Leader] of Letters), who popularized the *kiḷippāṭṭu* ("parrot song"), a genre in which the narrator is a parrot, a bee, a swan, and so on. His outstanding works are *Adhyātma Rāmāyaṇam*, *Bhāratam*, and *Bhāgavatam*, all based on Sanskrit originals yet powerfully re-created with masterly language craft.

While Vaiṣṇava works were proliferating in Malayalam, Śaiva movements swept the other three languages, Tamil, Kannada, and Telugu. In Tamil, the hymns of the Nāyaṉārs were arranged and anthologized for scriptural and recitative use by the 11th century. Another such consolidation of sacred materials was Cēkkiḻār's 12th-century *Toṇṭar Purāṇam*, or *Periyapurāṇam*, narrating in epic style the lives of the 63 great Śaiva saints and creating a tradition for all Śaivas, even in the Kannada and Telugu areas. The theology of the Siddhānta (Perfected Man) school of Śaivism was elaborated in Meykaṇṭār's *Civañāṇapōtam* (13th century).

The new Kannada genre of the 13th century

By the 12th century, a new Kannada genre, the *vacana* ("saying" or "prose poem"), had come into being with the Vīraśaiva saints. In the language of the people, the saints expressed their radical views on religion and society, rejected both Brahminical ritualism and Jaina ascetic world negation, called all men to the *anubhāva* ("experience") of God, and broke the bonds of caste, creed, and sexual difference. Five important poet-saints were Dāsimayya; Basava, a self-searching social reformer and a minister of the Jaina king Bijjaḷa; Allama Prabhu, the elder and metaphysical master of them all; Mahādēviyakka, a woman saint singing love poems to Śiva; and Cannabasava, a brilliant theologian of the movement, who elaborated the theory of "six stages" of mystic ascent for the devotee. The many-facetted lyrics written by the poet-saints were in the spoken dialects of Middle Kannada, yet they drew on archetypal human images as well as ancient pan-Indian symbology for their intense and searing expressions of *bhakti*. Inspired by these lyrics, Harihara, in the late 12th century, wrote some 120 *ragaḷe* (blank verse) biographies of the Śaiva saints, including the Vīraśaiva (or Liṅgāyat) and the earlier Tamil Nāyaṉārs. In the early 13th century, his disciple and nephew, Rāghavāṅka, wrote, in *ṣaṭpadis* (six-line stanzas), of the lives of saints, in well-structured works such as *Sōmanātha Carite* and *Siddharāma Caritra;* his most mature work is *Hariścandrakāvya*, an unequalled reworking of an ancient Job-like story of Hariścandra, who suffered every ordeal for his love of truth. The Vīraśaiva saints' lives and the *vacana* ("saying" or "prose poem") literature were codified in a masterpiece called *Śūnya Sampādane* ("The Achievement of Nothing"), consisting of dialogues interweaving the saints' *vacanas*, with the poet Allama Prabhu as the central figure.

Contemporary with the 13th-century Vīraśaiva saints were Telugu Śaiva poets such as Pālkuriki Sōmanātha, who composed the *Basavapurāṇam* employing popular metres and idiomatic Telugu. His *Paṇḍitārādhya Caritra* is a life of the Śaiva devotee Paṇḍitārādhya as well as a book of general knowledge including social customs, arts, crafts, and particularly music. His *Vṛṣādhipa Śatakam* consists of verses in Tamil, Kannada, Marathi, Sanskrit, and Telugu. This work was probably the first of the genre of *śatakas* ("centuries" of verses) literature, particularly popular in Telugu but also written in the other three languages as well as in Sanskrit (see above *Sanskrit: formative period* [*1200–400 BC*]).

Also of the 13th century is Āṇḍayya's *Kabbigara Kāva* ("The Poet's Defender"), in Kannada, a linguistic tour de force, eschewing unmodified Sanskrit forms; and Mallikārjuna's *Sūktisudhārṇava*, an excellent Kannada anthology of lyrics and passages. From 1240 to 1326, poets of Telugu produced over 100 verse renderings of the Sanskrit epic *Rāmāyaṇa* and many more in prose, the earliest being *Raṅganātha Rāmāyaṇa*, assigned to Gōna Buddhā Reḍḍi.

14th–19th century. The next age, from the 14th to the 16th century, is the great age of the Vijayanagar Empire. In this period, Kannada and Telugu were under the aegis of one dynasty and were also hospitable to the influence of neighbouring Muslim Bahmanī kingdoms. Śrīnātha was a 15th-century poet honoured in many courts for his scholarship, poetry, and polemics. He rendered Sanskrit poems and wrote *Haravilāsam* (Four Śaiva Tales); *Krīḍābhirāmam*, a charming, often vulgar account of social life in Warangal; and *Palanāṭi Vīra Caritra*, a popular ballad on a fratricidal war. Many erotic *cāṭus*, or stray epigrams, are also attributed to him. Bammera Pōtana, a great Śaiva devotee in life and poetry, unschooled yet a scholar, is widely known for his *Bhāgavatam*, a masterpiece that

is said to excel the original Sanskrit *Bhāgavata-Purāṇa.* Tāḷḷapāka Annāmācārya, son of a great family of scholars, fathered an exciting new genre of devotional song, all addressed to the god Śrī Veṅkaṭeśvara of Tirupati (a form of Vishnu). His *Saṅkīrtana Lakṣaṇam* is a collection of 32,000 songs in Sanskrit and Telugu, which made a significant contribution to Karnatic (southern Indian) musical technique.

The 16th century was an age of patronage by Vijayanagar kings, beginning with Kṛṣṇa Dēva Rāya, himself a poet versed in Sanskrit, Kannada, and Telugu. The *rāyala yugam* ("age of kings") was known for its courtly *prabandhas,* virtuoso poetic narratives by and for pandits (learned men). Among the most famous court poets were Piṅgaḷi Sūranna, whose verse novel, *Kalāpurṇōdayam* (1550)—a story full of surprises, magic, and changes of identity—is justly celebrated for its artistry; and Tenāli Rāmakṛṣṇa, known for his clownish pranks and humour, whose writings are the centre of a very popular cycle of tales in all four Dravidian languages.

During the 16th century and for the next few centuries, Telugu poets also flourished outside the Telugu country, especially in Tanjore (Thanjavūr) and Madurai, in Tamil country, and Pudukkoṭṭa and Mysore, in Kannada country. Their most important contribution was to native Kannada and Telugu dance drama on mythological themes, called *yakṣagāna.* The form is comparable to *kathākali* in the Malayalam area and to *terukkūttu* ("street drama") and *kuṟavañci* ("gypsy drama") in the Tamil area. The earliest Telugu *yakṣagāna* text is *Sugrīva Vijayam* (*c.* 1570), by Kandukuru Rudra Kavi; the earliest in Kannada is probably Śāntavīra Dēśika's *Saundarēśvara* (1678). The most celebrated of Kannada *yakṣagāna* dramatists is the versatile Pārti Subba, who flourished around 1800 and is known for his moving *Rāmāyaṇa* episodes and songs.

Contribution of poets to native dance drama

The 15th and 16th centuries produced some of the most popular classics in Kannada. Of these the greatest is Gadugu's *Kumāra Vyāsa,* or *Nāraṇappa*'s, 10 cantos of the *Mahābhārata;* recited in assemblies as well as in households, these are a continual delight, abounding in humour, passion, and memorable poetry. In *Prabhuliṅgalīle,* Cāmarasa made poetry out of the life of the poet-saint Allama. The *Jaimini Bhārata* and the many versions of *Rāmāyaṇa* episodes (especially Sītā's abandonment in the forest) written by the distinguished Śaiva epic poet Lakṣmīśa are known for their melodious verses and moving scenes. Ratnākaravarṇi's *Bharateśa Vaibhava* is a great Jaina story, tersely told in a Kannada song metre and celebrated for its depiction of many *rasas* ("moods"), especially the erotic.

Kannada Vaiṣṇava *dāsas* ("servants [of God]") wrote in a song genre called *pada,* parallel and often indebted to the Vīraśaiva *vacanas* ("sayings" or "prose poems"). Purandaradāsa, a rich 16th-century merchant turned mendicant, saint, and poet, composed *bhakti* (devotional) songs on Viṭṭhala (a manifestation of the god Vishnu), criticizing divisions of caste and class and calling on the mercy of God. His *padas* and *kīrtanas* ("lauds") are also landmarks in Karnatic music. Karnatic music. Kanakadāsa, his contemporary and a shepherd by birth, wrote *padas* and longer popular works. *Dāsa* songs are part of the repertory of all South Indian musicians.

The folk *tripadi* ("three-line verse") of Sarvajña (1700?) is a household word for wit and wisdom, like the *Kuṟaḷ* in Tamil (see above *Eighteen Ethical Works*) and the "century" of four-line verses in Telugu by Vēmana (15th century). The moral, social, satiric, and wise proverb-like aphorisms of Vēmana and Sarvajña are widely quoted by pundit and layman alike. Equally popular in the Malayalam region is the 18th-century folk poet of *tuḷḷals* (a song-dance form), Kuñcan Nampiyār, unparalleled for his wit and exuberance, his satiric sketches of caste types, his versions of Sanskrit *Purāṇa* narratives projected on the backdrop of Kerala, and his humorous renderings even of mythic characters.

The 17th and 18th centuries also saw Tamil court poetry—*Purāṇas,* translations from the Sanskrit, and praise poems, known more for their learning and imitative character than for their genius. This was also a period of many schisms and the founding of monasteries in Śaivism and Vaiṣṇavism, which led to many sectarian and polemic works. Muslims and Christians also wrote epics in the Hindu *Purāṇa* style; for example, Umaṟu-p-pulavar's 17th-century *Cīṟā-p-purāṇam,* on the life of the prophet Muḥammad, and Father Beschi's *Tēmpāvaṇi,* on the life of St. Joseph, with echoes from both Kampaṉ and the 16th-century Italian poet Torquato Tasso.

Probably the most impressive Tamil poetry of this period is that of Arunakiriv's learned and melodious *Tiruppukaḷ* (praise of Munikaṉ) and of the Cittars, eclectic mystics known for their radical, fierce folk songs and common-speech style. Tāyumāṇavar (18th century) and Paṭṭiṉattār (and later, in the 19th century, Rāmaliṅkar) are poets of unconditioned love, self-search, and rejection of corrupt society.

The 17th and 18th centuries are also periods of datable folk expression, which include many *tiruviḷaiyāṭal* ("stories of God's sport") *purāṇas;* temple tales (about miracles that took place in the temple); *kuṟavañci* (*i.e.,* "gypsy," a kind of musical dance drama); *paḷḷus* (plays about village agricultural life); realistic *noṇṭi-nāṭakams* ("dramas of the lame"), in which a Hindu temple god cures lameness; *kummi* songs sung by young girls, clapping as they dance round and round; and *ammāṇai* ballads. Noteworthy historical ballads are *Kaṭṭa Pommaṉ,* about a chieftain who revolted against the British, and *Tēciṅku-rācaṉ Katai,* about the prince of Gingi and his Muslim friend. Malayalam *āṭṭakkatha,* the literature associated with *kathākali,* the complex traditional dance drama, was also written during this period. Royal poets such as Kōṭṭayattu Tampurān, in the 17th century, and Kārttika Tiruṇal, in the 18th, wrote *āṭṭakkathās.* (A.K.R.)

INDO-ARYAN LITERATURES: 12TH–18TH CENTURY

It is difficult to pinpoint the time when the Indo-Aryan dialects first became identifiable as languages. Around the 10th century AD, Sanskrit was still the language of high culture and serious literature, as well as the language of ritual. The spoken language, however, had continued to develop, and at the turn of the millennium there began to appear, at different times during the subsequent two or three centuries, the languages now known as the regional languages of the subcontinent: Hindi, Bengali, Kashmiri, Punjabi, Rajasthani, Marathi, Gujarati, Oriya, Sindhi (which did not develop an appreciable literature), and Assamese; Urdu did not develop until much later (see below *Islāmic literatures: 11th–19th century*).

Regional languages

The literatures in their early stages show three characteristics: first, a debt to Sanskrit that can be seen in their use of Sanskrit lexicon and imagery, in their use of myth and story preserved in that refined language, and frequently in their conformity to ideals and values put forward in Sanskrit texts of poetics and philosophy; second, a less obvious debt to their immediate Apabhraṃśa past (dialects that are immediate predecessors of the modern Indo-Aryan vernaculars); third, regional peculiarities.

The narratives in the early stages of the development of the languages are most often mythological tales drawn from the epics and *Purāṇas* of classical Hindu tradition (see above *Sanskrit, Pāli, and Prākrit literatures: 1400 BC–AD 1200*), though in later times, in the 17th and 18th centuries, secular romances and heroic tales were also treated in narrative poems. Although the themes of the narratives are based on *Purāṇa* tales, often they include materials peculiar to the area in which the narrative was written.

In addition to themes, regional literatures frequently borrowed forms from the Sanskrit; for example, the *Rāmāyaṇa* appears in a 16th-century Hindi version by Tulsīdās, called the *Rāmcaritmānas* ("Lake of Rāma's Deeds"), which has the same form, though a different emphasis, as the Sanskrit poem. The stylized conventions and imagery of Sanskrit court poetry also appear, though here, too, with different emphasis; for example, in the work of the 15th-century Maithili (Eastern Hindi) lyric poet Vidyāpati. Even the somewhat abstruse rhetorical speculations of the Sanskritic poetic schools of analysis were used as formulas for the production of 17th-century Hindi court poetry; the

Rasikapriyā ("Beloved of the Connoisseur") of Keśavadāsa is a good example of this kind of tour de force.

There are other characteristics common to the regional literatures, some of which come not from Sanskrit but most likely from the Apabhraṃśa. There are two poetic forms, for example, that are found in many northern Indian languages: the *bārah-māsā* ("twelve months"), in which 12 beauties of a girl or 12 attributes of a deity might be extolled by relating them to the characteristics of each month of the year; and the *caūtīs* ("thirty-four"), in which the 34 consonants of the northern Indian Devanāgarī alphabet are used as the initial letters of a poem of 34 lines or stanzas, describing 34 joys of love, 34 attributes, and so on.

Finally, there are common characteristics that may have come either through Apabhraṃśa or through the transmission of stories and texts from one language to another. The stories of Gopi-candra, the cult hero of the Nātha Yogī sect, a school of mendicant *sannyāsins*, were known from Bengal to the Punjab even in the early period. And the story of the Rājput heroine Padmāvatī, originally a romance, was beautifully recorded, with a Ṣūfī (mystic) twist, by the 16th-century Muslim Hindi poet Malik Muḥammad Jāyasī and later by the 17th-century Bengali Muslim poet Ālāol. From the late 13th through the 17th century, bhakti (devotional) poetry took hold in one region after another in northern and eastern India. Beginning with the *Jñāneśvarī*, a Marathi verse commentary on the *Bhagavadgītā* written by Jñāneśvara (Jñānadeva) in the late 13th century, the devotional movement spread through Mahārāshtra, in the works of the poet-saints Nāmdev and Tukārām; through Rājasthān, where it is represented by the works of Mīrā Bāī; through northern India, in the poetry of Tulsīdās, Sūrdās, Kabīr, and others; through Mithilā, in the work of the great poet Vidyāpati; and into Bengal, where Caṇḍīdās and others sang of their love of God. Because of the *bhakti* movement, beautiful lyric poetry and passionate devotional song were created; and in some cases, as in Bengal, serious philosophical works and biographies were written for the first time in a regional language rather than in Sanskrit. The languages and their literatures gained strength as mediums of self-expression as well as exposition. And, although there is much Sanskrit imagery and expression in the poetry and song, as well as similarities to Sanskrit textual models, its basic character is not Sanskritic: true to the nature of any spoken, everyday language, it is more vital than polished, more vivid than refined.

Spread of *bhakti* poetry

One more historical generality can be stated regarding regional Indian literature before considering the characteristics peculiar to the several "Indian literatures." In all of the early literatures, writing was lyrical, narrative, or didactic, entirely in verse, and all in some way related to religion or love or both. In the 16th century, prose texts, such as the Assamese histories known as the *buranji* texts, began to appear.

Hindi. What is commonly spoken of as Hindu is actually a range of languages, from Maithili in the east to Rajasthani in the west. The first major work in Hindi is the 12th-century epic poem *Pṛthvīrāj Rāsau*, by Chand Bardaī of Lahore, which recounts the feats of Pṛthvīrāj, the last Hindu king of Delhi before the Islāmic invasions. The work evolved from the bardic tradition maintained at the courts of the Rājputs. Noteworthy also is the poetry of the Persian poet Amīr Khosrow, who wrote in the Awadhi dialect. Most of the literature in Hindi is religious in inspiration; in the late 15th and early 16th centuries, the reform-minded Kabīr, for example, wrote sturdy short poems in which he sought to reconcile Islām and Hinduism.

The most celebrated author in Hindi is Tulsīdās of Rājāpur (died 1623), a Brahmin who renounced the world early in life and spent his days in Benares (Vārānasi) as a religious devotee. He wrote much, mostly in Awadhi, and focussed Hinduism on the worship of Rāma. His most important work is the *Rāmcaritmānas* ("Sacred Lake of the Acts of Rāma"), which is based on the Sanskrit *Rāmāyaṇa*. More than any other work it has become a Hindu sacred text for the Hindi-speaking area and annually has been staged in the popular Rām Līlā festival.

Outstanding among the followers of Vallabha, philosopher and *bhakti* ("devotion") advocate of the Middle Ages, is the blind poet Sūrdās (died 1563), who composed countless *bhajans* (chants) in praise of Krishna and Rādhā, which are collected in the *Sūrsāgar* ("Ocean of Sūrdās"). While many of the *bhakti* poets were of modest origin, an exception was Mīrā Bāī, a princess of Jodhpur, who wrote her famous lyrics both in Hindi and Gujarati; the quality of her poetry, still very popular, is not as high, however, as that of Sūrdās. Significant also is the religious epic *Padmāvatī* by Jāyasī, a Muslim from former Oudh state. Written in Awadhi (*c.* 1540), the epic is composed according to the conventions of Sanskrit poetics.

The 18th century saw the beginning of a gradual transformation from the older forms of religious lyric and epic to new literary forms influenced by Western models that began to be known. The new trends reached their pinnacle in the work of Prem Chand (died 1936), whose novels (especially *Godān*) and short stories depict common rural life; and in the work of Harishchandra of Benares (died 1885), honoured as Bhāratendu (Moon of India), who wrote in the Braj Bhasa dialect.

Bengali. While developments in Bengali literature began somewhat earlier, they followed the same general course as those in Hindi. The oldest documents are Buddhist didactic texts, called *caryā-padas* ("lines on proper practice"), which have been dated to the 10th and 11th centuries and are the oldest testimony to literature in any Indo-Aryan language.

The three genres of Bengali poetry

Bengali poetry, including poetry by Bengalis in other dialects, is largely written in three distinct genres. It is certain that well before the 15th century there existed texts in a typically Bengali genre called *maṅgal-kāvya* ("poetry of an auspicious happening"), which consists of eulogies of gods and goddesses; such poetry is likely to have had a considerable history in oral transmission before it was committed to writing. A good example of an orally transmitted *maṅgal* poem is the *Caṇḍī-maṅgal* ("Poem of the Goddess Caṇḍī"), by Mukundarāma, which was put into written form in the latter part of the 15th century. *Maṅgal* poetry remained a favourite genre well into the 18th century, when Bhārat-candra wrote the *Annadā-maṅgal* ("Maṅgal of the Goddess Annadā [the Giver of Food]"), a witty and sophisticated poem that bears little resemblance to its more rustic forebears. Despite this popularity, it is the devotional lyrics to the divine pair Krishna and Rādhā that are still known and sung today in Bengal, and these lyrics are the gems of medieval Bengali literature.

Poems of the second genre, the *mahākāvya* ("great poem," but not to be confused with the Sanskrit *mahākāvya* genre), are based mainly on the Sanskrit models of the *Mahābhārata*, *Rāmāyaṇa*, and *Purāṇas*. Kṛttibās Ojhā (late 14th century) stands at the beginning of this literature; he wrote a version of the *Rāmāyaṇa* that often differs from the Sanskrit original, for he includes many local legends and places the setting in Bengal. Kavīndrā (died 1525) wrote on the *Mahābhārata* theme, as did Kāsiram Dās in the 17th century.

The third genre, *padāvalī* ("string of verse") songs, is also found elsewhere; inspired by the religious *bhakti* movement, the songs resemble the devotional poetry of the Nāyaṇārs and Āḻvārs in Tamil. It was such poetry that established Bengali as a significant literary language. The earliest work in what may be considered a distinctively Bengali style is the *Śrīkṛṣṇa-kīrtana* ("Praise of the Lord Krishna"), a long *padāvalī* poem by Caṇḍīdās, which is dated to the early 15th century. In it the poet praises the virtues and celebrates the loves of Krishna, a theme that had remained popular in Bengal ever since its first glorification by the Bengali Sanskrit poet Jayadeva, who composed his *Gītagovinda* ("The Cowherd Song") in the 12th century. *Padāvalī* songs describe and glorify all phases of Krishna's love for the cowherds' wives (especially Rādhā, who later became a goddess), and it is love poetry before it is religious poetry. After the great Bengali mystic and saint Caitanya (died 1533), love *is* religion, and the erotic is inspirited with religious fervour. The great flowering of this poetry occurred in the 16th and 17th centuries.

Religious edification took the forms not only of *maṅgals*

and *padāvalī*s but also of biography (more like hagiography) and philosophy. Important in that style is the long hagiography *Caitanya-caritāmṛta* ("Elixir of the Life of Caitanya"), by the 16th-century author Kṛṣṇadās.

While most of the literature is Hindu in theme and inspiration, there arose a secular Bengali literature among Bengali Muslims. One of the outstanding Muslim poets is Ālāol, author of the *Padmāvatī* (*c.* 1648), which was written after the poem of the same name by the Hindi poet Jāyasī.

Assamese. The earliest text in a language that is incontestably Assamese is the *Prahlāda-caritra* of Mena Sarasvati (13th century); in a heavily Sanskritized style it tells the story, from the *Viṣṇu-Purāṇa,* of how the mythical king Prahlāda's faith and devotion to Vishnu saved him from destruction and restored the moral order. The first great Assamese poet was Kavirāja Mādhava Kandalī (14th century), who translated the Sanskrit *Rāmāyaṇa* and wrote *Devajit,* a narrative on the god Krishna. In Assamese, too, the *bhakti* movement brought with it a great literary upsurge; the most famous Assamese poet of the period was Śaṅkaradeva (died 1568), whose 27 works of poetry and devotion are alive today and who inspired such poets as Mādhava-devi to write lyrics of great beauty. Peculiar to Assamese literature are the *buranjī*s, chronicles written in a prose tradition brought to Assam by the Ahoms of Burma. These date in Assamese from the 16th century, while in the Ahom language they are much earlier.

Oriya. *Mādaḷā-pāñji* ("The Drum Chronicle") texts in Oriya, the chronicles of the great temple of Jagannātha in Puri, date from the 12th century. They are in prose, and as such they represent the earliest prose in a regional Indo-Aryan language, although they cannot be said to be literary texts. The 14th century was productive for Oriya literature. Dating from this period are the anonymous *Kalasa-cautīśa,* which tells in 34 verses the story of the marriage of the god Śiva and the mountain goddess Pārvatī, and the famous *Caṇḍī-purāṇa* of Saraladāsa. But the *bhakti* period was once again the most stimulating one; the best known medieval Oriya poet is Jagannātha Dās (whose name means Servant of Jagannātha), a 16th-century disciple of the Bengali Vaiṣṇava saint Caitanya, who spent the better part of his life in Puri. Among the many works of Jagannātha Dās is a version of the Sanskrit *Bhāgavata-Purāṇa* that is still popular in Orissa today.

Marathi. With Bengali, Marathi is the oldest of the regional literatures in Indo-Aryan, dating from about AD 1000. In the 13th century, two Brahminical sects arose, the Mahānubhāva and the Varakari Panth, both of which put forth vast quantities of literature. The latter sect was perhaps the more productive, for it became associated with *bhakti,* when that movement stirred Mahārāshtra in the early 14th century, and particularly with the popular cult of Viṭṭhoba at Pandharpur. It was out of this tradition that the great names of early Marathi literature came: Jñāneśvara, in the 13th century; Nāmdev, his younger contemporary, some of whose devotional songs are included in the holy book of the Sikhs, the *Ādi Granth;* and the 16th-century writer Eknāth, whose most famous work is a Marathi version of the 11th book of the *Bhāgavata-Purāṇa.* Among the *bhakti* poets of Mahārāshtra the most famous is Tukārām, who wrote in the 16th century. A unique contribution of Marathi is the tradition of *povāḍā*s, heroic stories popular among a martial people. There is no way of dating the earliest of these; but the literary tradition is particularly vital at the time of Śivajī, the great military leader of Mahārāshtra (born 1630), who led his armies against the might of the Mughal emperor Aurangzeb.

Tradition of *povāḍā*s

Gujarati. The oldest examples of Gujarati date from the writings of the 12th-century Jaina scholar and saint Hemacandra. The language had fully developed by the late 12th century. There are works extant from the middle of the 14th century, didactic texts written in prose by Jaina monks; one such is the *Bālāvabodha* ("Instructions to the Young"), by Taruṇa-prabha. A non-Jaina text from the same period is the *Vasanta-vilāsa* ("The Joys of Spring"). The two Gujarati *bhakti* poets, both of the 15th century, are Narasiṃha Mahatā (or Mehtā) and Bhālaṇa (or Puruṣottama Mahārāja); the latter cast the 10th book of the *Bhāgavata-Purāṇa* into short lyrics. By far the most famous of the *bhakti* poets is the woman saint Mīrā Bāī, who lived in the first half of the 16th century. Mīrā, though married, thought of Krishna as her true husband, and the lyrics telling of her relationship with her god and lover are among the warmest and most movingly personal in any Indian literature. One of the best known of the non-*bhakti* Gujarati poets is Premānanda Bhaṭṭa (16th century), who wrote narrative poems based on *Purāṇa*-like tales; although his themes were conventional, his characters were real and vital, and he infused new life into the literature of his language.

Punjabi. Punjabi developed a literature later than most of the other regional languages of the subcontinent; and some of the early writings, such as those of the first Sikh Gurū, Nānak (late 15th and early 16th centuries), are in Old Hindi rather than true Punjabi. The first work identifiable as Punjabi is the *Janam-sākhī,* a 16th-century biography of Gurū Nānak by Bala. In 1604, Arjun, the fifth Gurū of the Sikhs, collected the poems of Nānak and others into what is certainly the most famous book to originate in the Punjab (though its language is not entirely Punjabi), the *Ādi Granth* ("First Book"). Writing that is not merely incidentally Punjabi began in the 17th century and is almost entirely by Muslims. Between 1616 and 1666, a writer named ʿAbdullāh, for example, composed a major work called *Bāra Anva* ("Twelve Topics"), which is a treatise on Islām in 9,000 couplets. Muslim Ṣūfīs, such as Bullhē Shāh (died 1758), also contributed many devotional lyrics, and Ṣūfī Islām can be said to have been the main stimulus to Punjabi literature in the medieval period. There are also many romances in the language (as in Rajasthani) which, being oral literature, are undatable.

Kashmiri. The hitherto commonly accepted period of Old Kashmiri is 1200–1500; but in fact the earliest example of the language is found in 94 four-line stanzas embedded in the Sanskrit philosophical work *Mahānaya-prakāśa* ("Illumination of the Highest Attainment"), which some scholars now date as late as the 15th century. As is true for Gujarati, the most famous poets of Kashmiri in this period are women. Lallā (14th century) wrote poems about the god Śiva; and Hubb Khātun (16th century) and especially Arani-mal (18th century) are famous for their hauntingly beautiful love lyrics. Despite these outstanding poets in Kashmiri, the great literary language of Kashmir in the medieval period was Persian, which was encouraged by many rulers of the country, such as Zayn-ul-ʿĀbidīn, in whose 15th-century court were many scholars and poets writing in both the Kashmiri and Persian languages.

(E.C.D.)

ISLĀMIC LITERATURES: 11TH–19TH CENTURY

The adventure of Islām in India began in the 8th century with the conquest of Sind (the extreme western province), but it was only in the 11th and 12th centuries that Muslim literary and cultural traditions reached the Indian heartland. Then, in the 13th century, refugee noblemen, soldiers, and men of letters from Iran and Central Asia came pouring into India. Although the causes changed, the attraction of India as a place of refuge and gracious patronage did not decline for several subsequent centuries. At the same time Muslim soldier-adventurers continued with their conquests, joining hands with their non-Muslim Indian counterparts in many instances, establishing minor or major kingdoms all over the subcontinent. The political map of India remained very much in flux—except for a brief period during the reign of Akbar—until Queen Victoria declared herself empress of India in 1858. The period of Muslim influence thus extends over 800 years.

At the time of the spread of Muslim power and culture in India, Sanskrit was the chief language of Hindu cultural, learned, and religious expression, while Buddhism and Jainism had lent their prestige and patronage to various Prākrits. The progress of and developments in these literatures remained unaffected by the advent of Islām in India. The emergence of the new Indo-Aryan languages out of the Prākrit and Apabhraṃśa stages of Sanskrit, however, was furthered by the newcomers, who preferred these regional languages over Sanskrit and encouraged the development

Effect of Islām on Sanskrit and Prākrit literatures

of popular regional literatures. The conversion to Islām of a large number of indigenous people enhanced these developments. Thus, the vehicles of literary expression used by those professing Islām in India were regional dialects and languages, both Indo-Aryan and Indo-Iranian, such as Braj, Awadhi, Sindhi, Baluchi, Urdu, Dakhini, and Bengali, as well as the foreign Arabic, Turkish, and Persian spoken by the Muslim immigrants and conquerors. Of these, only Persian and Urdu require detailed consideration; the others will be discussed only briefly.

Arabic. Arabic was the language of the conquerors of Sind. But it enjoyed more permanent prestige as the language of the Qur'ān, the sacred book of Islām; as such it was extensively used for religious scholarship during the medieval period. Even as late as the 18th century, Shāh Walī Allāh, the greatest theologian to have lived in India, wrote his most important treatises in Arabic. Arabic was also used early for historiography and for making Indian scientific books available to the Middle East in translation. One does not find, however, much in the way of significant Arabic belles lettres in India.

Turkish. Although the earliest Muslim conquerors in northern India were Turks, their language was Persian. It was only during the reigns of Bābur and his son Humāyūn (1526–56) that Turkish flourished for a while as a medium of learned expression. Bābur himself was the foremost contributor. Although his memoirs are better known, he also left a volume of verses of considerable merit.

Regional languages. The literatures of the Indo-Iranian languages of Baluchi and Pashto are exclusively creations of Muslim writers. In the Indo-Aryan languages of Kashmiri, Sindhi, and Punjabi, Muslims were the most influential contributors; the names of Lallā (14th century) for Kashmiri, Shāh ʿAbd-ul-Laṭīf (17th–18th century) for Sindhi, and Wāriš Shāh (18th century) for Punjabi exemplify that fact. Muslim chieftains gave impetus to the growth of Bengali literature through their patronage of writers and through their efforts to have Sanskrit classics translated into Bengali. There are also many famous Muslim names during the medieval period of Bengali literature, such as Dawlat Qāzī and Ālāol in the 17th century. In the heartland of northern India, notable contributions were made by Muslims to Hindi literatures in the Braj and Awadhi dialects. Malik Muḥammad Jāyasī, Raḥīm, and Manjhan (all 16th century) and ʿUsman (17th century) are some of the important names. In the 16th, 17th, and 18th centuries in India there was a tremendous production of mystic (Ṣūfī and *bhakti*) poetry in all of the important dialects and languages. It was a period of great mystic, syncretic movements, and the Muslim contribution in the form of love narratives and lyrics was considerable. Quite often metres, motifs, and assorted rhetorical features of Persian *mašnavīs* and *ghazals* (see below *Urdu*) were used in a new medium. Moreover, interaction and assimilation took place between the Muslim Ṣūfī traditions, thought, and practices and the Indian *bhakti* schools. Muslim *bhakti* poets either expressed Ṣūfī ideas, which were close to monotheistic orthodoxy as well as to the doctrines of Indian saints Kabīr and Nānak, in the Indian dialects through narrative poems modelled on Persian *mašnavīs* or chose the path of ecstasy and became devotees of Krishna (which was still close to the more orthodox forms of Ṣūfīsm). None of them followed the devotional style of Tulsīdās, their contemporary and a devotee of Rāma.

It was, however, in Persian and Urdu that Muslim men of letters made the greatest contributions—contributions that led in the former case to the establishment of an "Indian" school of Persian poetry and influenced profoundly the development of poetry in Afghanistan and Tadzhikistan and, in the latter case, led to the emergence of a unique pan-Indian language and literature in Urdu.

Persian. Maḥmūd of Ghazna, with whom the chain of Muslim conquests in northern India began, was also the patron of Ferdowsī, one of the greatest of Persian poets. The later conquerors admired literature no less. Since the language of all of them was Persian, the growth of Persian literature in India kept pace with its conquest by the Muslims.

Masʿūd Saʿd Salmān (born 1046 in Lahore), who later became the governor of Jullundhur, was the first noteworthy person of Indian origin to have written poetry in Persian. The first truly great poet was Amīr Khosrow, who wrote in the 13th and 14th centuries. Of Turkish descent, born in the district of Etah in northern India, Khosrow was connected with royal courts all his life, even after 1272, when he became a disciple of the great mystic Niẓām-ud-Dīn Awliyā. He wrote five books of poems, or *dīvāns*, composed of *ghazals* (see below *Urdu*), panegyrics and several *mašnavīs*—altogether some 200,000 couplets. In poetry, his innovative spirit displayed itself in *waṣf-nigārī*—that is, descriptions of natural objects in short poems, which Khosrow incorporated within longer ones. His keenness of observation is also evident in his use of local fauna and flora as poetic images. Khosrow's distinction lies not so much in the fact that he is an innovator, however, as in the fact that he is equally superb in narrative poetry, panegyrics, and lyrics. The range of his popularity and influence can best be gauged by the fact that, in northern Indian folk literature, one comes across numerous songs and riddles consistently attributed to Amīr Khosrow.

The poems of Amīr Khosrow

In the centuries that followed Khosrow, until the end of the Islāmic period, India contributed to Persian literature in two ways: first, through the production of dictionaries that helped to standardize the language and consolidate its vocabulary; second, through the development of the so-called Indian style of Persian poetry.

It is generally agreed that this Indian style, *sabk-e hindī*, did not originate within the geographic confines of India, though it reached its most sublime form there at the hands of poets who either were born in India or spent their most productive years at various Indian courts. Some of the characteristics of the style are (in the words of one modern critic) the emphasis on

> parallel statement . . . ; on complex conceit like that of the seventeenth century English "metaphysical" poets, arising out of economy of expression and telescoping into a single image a variety of emotional states; on "cerebral" artifice in pushing familiar images to unfamiliar and unexpected lengths; and on the creation of a synthetic poetic diction in which a whole phrase constitutes a single image.

The keen observation of daily life that is also characteristic of Indian Persian poetry could have been inspired by the traditions of classical Sanskrit poetry, with which these poets must have been familiar through the extensive translations done during the reign of the Mughals.

The century (1556–1657) of the reigns of Akbar, Jahāngīr, and Shāh Jahān was the most glorious period for Persian poetry in India, though, except for Fayẓī, all of the important poets were immigrants from Persia who found relief from religious and political persecution as well as generous patronage at the hands of the great Mughals and the lesser kings of southern India. The great men of letters of that period were ʿUrfī, Ṭālib Āmulī, Naẓīrī, Ẓuhūrī, Kalīm, and Ṣāʾib.

The greatest poet of the Indian style, however, was ʿAbdul Qādir Bēdil, born in 1644 in Patna, of Uzbek descent. He came early under the influence of the Ṣūfīs, refused to be attached to any court, and travelled widely throughout India during his long life. Bēdil's 16 books of poetry contain nearly 147,000 verses and include several *mašnavīs*. Though ignored by the Iranians, Bēdil's poetry had an impact on Tadzhik and Uzbek literatures, and its influence is still evident in Afghanistan. A poet of great virtuosity and philosophic bent, he was well acquainted with Indian religions and philosophy. His anti-feudal views and his critical and skeptical attitude toward all kinds of dogma make his poetry relevant even today. His style is difficult, his metaphors and syntax quite complex (though the language itself is quite simple); and yet, as a modern critic puts it, "the intensity of his subjective assessment is so acute and factual, and his metaphysical experience so intense, that genuine poetry emerges in all its splendour."

The works of ʿAbdul Qādir Bēdil

Urdu. Earlier varieties of Urdu, variously known as Gujari, Hindawi, and Dakhani, show more affinity with eastern Punjabi and Haryani than with Khari Boli, which provides the grammatical structure of standard modern Urdu. The reasons for putting together the literary products of these dialects, forming a continuous tradition with

those in Urdu, are as follows: first, they share a common milieu, consisting of Ṣūfī and Muslim court culture, increasingly dominated by the life and values of the urban elite; second, they display wholesale acceptance of Perso-Arabic literary traditions, including genres, metres, and rhetoric; third, they show an increasing acceptance of Perso-Arabic grammatical devices and vocabulary; and fourth, they tend to prefer Perso-Arabic forms over indigenous forms for learned usage.

Apart from themes and metaphysics, the influence of Ṣūfī hospices and royal courts can be seen in two practices that were essential to the development of Urdu poetry (and also unique to the Urdu milieu in the medieval period) and that still exist in modified forms. First, Urdu poets generally chose an *ustād,* or master, just as a Ṣūfī novice chose a *murshid,* or preceptor, and one's poetic genealogy was always a matter of much pride. Second, poets read poetry in private or semiprivate gatherings, called *mushā'irah,* which displayed hierarchies, status consciousness, and rivalries reminiscent of royal courts.

Urdu literature began to develop in the 16th century, in and around the courts of the Quṭb Shāhī and 'Ādil Shāhī, kings of Golconda and Bījāpur in the Deccan (central India). In the later part of the 17th century, Aurangābād became the centre of Urdu literary activities. There was much movement of the literati and the elite between Delhi and Aurangābād, and it needed only the genius of Walī Aurangābādī, in the early 18th century, to bridge the linguistic gap between Delhi and the Deccan and to persuade the poets of Delhi to take writing in Urdu seriously. In the 18th century, with the migration of poets from Delhi, Lucknow became another important centre of Urdu poetry, though Delhi never lost its prominence.

The first three centuries are dominated by poetry. Urdu prose truly began only in the 19th century, with translations of Persian *dāstāns,* books prepared at the Delhi College and the Fort William College at Calcutta, and later with the writers of the Aligarh movement.

To focus on essential matters, the discussion that follows forgoes a chronological account of the poetry, concentrating instead on characteristics of particular genres and the achievements of the most significant of their practitioners up to 1857. There is one poet, however, who cannot be described as a practitioner of the classical Perso-Arabic traditions adopted by his fellow poets. Naẓīr Akbarābādī, who wrote in the late 18th and early 19th centuries, was a poet of consummate skill who chose to display it in short poems (in various forms) written in the language of popular speech as well as of literature. His themes show similar eclecticism. In his voluminous body of work, there are poems on such diverse topics as popular festivals, the seasons, the vanities of life, erotic pleasures and pursuits, dancing bears, and niggardly merchants. He is a master of the telling detail that immediately brings any event to life. Generally ignored by elitist poets and literary chroniclers of his time, Naẓīr has gained increasing respect and recognition as the first and best poet of the people.

The works of Naẓīr Akbarābādī

Qasīdahs. *Qasīdah*s are poems written with a "purpose"—the purpose being worldly gain, in the case of poems praising kings and noblemen, or benefit in the afterworld, in the case of poems praising God, the prophet Muḥammad, and other holy personages. These panegyrics are generally overly long and are written in a highly ornate and hyperbolic style, the poets vying to display their prowess by using as many rhymes and discovering as many associative themes as possible. Because of their style and language they are of special interest to lexicographers. Not much scholarly work has been done on the *qasīdah*s written in the Deccan, but in northern India a number of poets are regarded highly for their achievements in this genre: in the 18th century, Sawdā and Inshā', and in the 19th, Żawq and Ghālib.

Ḥaju and shahr-āshūb. Less ornate, if not less elaborate, and more edifying are the *ḥaju* (derogatory verses, personal and otherwise) and the *shahr-āshūb* (poems lamenting the decline or destruction of a city). They provide useful information about the mores and morals of the period from the 18th to mid-19th century and truly depict the problems facing the society at large. The poems are not formally restricted to any particular metre or stanza pattern. Sawdā again is one of the more famous names.

Marsiyeh. *Marsiyeh* means "elegy," but in Urdu literature it generally means an elegy on the travails of the family and kinsmen of Ḥusayn (grandson of Muḥammad) and their martyrdom in the field of Karbalā, Iraq. These elegies and other lamentatory verses were read at public gatherings, especially during the month of Muḥarram. Although a large number of *marsiyeh*s were written in the Deccan and at Delhi, it was in Lucknow, with the patronage of Shī'ite elite and royalty, that *marsiyeh*s gained the tenor and magnitude of epic poetry. The two great masters of that 19th-century period were Mīr Anīs and Mīrzā Dabīr, who together established *musaddas* (a six-line stanza with an *aaaa bb* rhyme scheme) as the preferred form for *marsiyeh*s and added several new topics and details to the ranks of associated themes, thus carrying the form beyond a simple lament. An interesting aspect of these elegies is that, although the scene and personae are Arab, there is no attempt at verisimilitude: Arab gallants and maidens speak and gesture like the elites of Lucknow. Perhaps this added to the pathos and effectiveness of the poems at public readings.

Urdu lamentatory elegies

Maśnavī. *Maśnavī* was the preferred genre for all descriptive and narrative purposes, for it allows the most freedom (only the lines of each couplet must rhyme). In the Deccan, all major poets wrote at least one long *maśnavī,* but lack of knowledge of the dialect has prevented their full appreciation. Thus, the more famous *maśnavī*s are by later poets of Delhi and Lucknow, such as Mīr, Mīr Ḥasan, Dayā Shankar Nasīm, and Mīrzā Shawq. The topics of descriptive *maśnavī*s range from mundane events of life, hunting trips of kings, and the vagaries of nature's seasons to autobiographical discourses. Narrative *maśnavī*s are considerably longer, running into hundreds of couplets. In the Deccan several poets wrote abridged versions of Persian *maśnavī*s, but many others wrote original compositions utilizing Indian romances as well as the better known Persian and Arabic ones. Apart from the names of the protagonists in the *maśnavī*s inspired by Persian and Arabic poems, all else is always local; the landscape, cityscape, processions, customs and rituals, social values and taboos, even the physical characteristics of the people are totally Indian, though dominantly Muslim and feudal. Despite their length, these narratives gained much popularity and, at least in northern India, were often read in public places, in much the same way as storytellers told stories. The *maśnavī* form was also used by some of the Hindi Ṣūfī poets.

Ghazal. For the most part, the history of Urdu poetry in India is the story of Urdu *ghazal,* which has been the favourite of both poets and their audiences in every period. A short lyric, with prosodic requirements of both metre and rhyme, *ghazal* demands great skill and thought from the poet, for its couplet must be a complete semantic entity and fully express a whole, well-integrated poetic experience. Favourite themes are erotic love, Ṣūfī love, and metaphysics. Naturally, Urdu poets began by closely imitating, often even plagiarizing, Persian masters, but later on they spoke in a more authentic voice. They continued, however, to employ a vocabulary of love that owed almost everything to Persian and shared very little with the traditions of lyrical poetry in other Indian languages. For example, with few exceptions, the lover is always masculine; expression of love is never made by a woman. Unique, too, is the use of masculine grammatical forms and imagery for the beloved, even when, in every other way, the poem is clearly celebrating heterosexual love. This peculiarity, as well as other traditions borrowed from Persian masters, gives a *ghazal* couplet a tremendously wide range of interpretations. It is amazing indeed what a master poet can condense into one terse couplet.

Characteristics of the *ghazal*

The two greatest *ghazal* writers in Urdu are Mīr Taqī Mīr, in the 18th century, and Mīrzā Asadullāh Khān Ghālib, in the 19th. They are in some ways diametrical opposites. The first prefers either very long metres or very short, employs a simple, non-Persianized language, and restricts himself to affairs of the heart. The other writes in metres of moderate length, uses a highly Persianized vo-

cabulary, and ranges wide in ideas. Mīr speaks of passion and pathos; Ghālib betrays a skeptic's mind and leaves nothing unquestioned, not even his feelings. But both have left indelible marks on the ideas and emotions of succeeding generations. Ghālib wrote poetry in Persian as well as Urdu and also published a couple of volumes of letters in Urdu that helped usher in modern prose. In many ways he bridges the gap separating the medieval sensibility from the modern. The contemporary mind, however, is also moved by the authentic passion of Mīr, idolizing him for the sublimity of his concept of love and for his personal integrity. The poems of Ghālib and Mīr represent the best of the Urdu *ghazal;* and the Urdu *ghazal,* as an anonymous wit has remarked, is the Muslims' greatest gift to India, after the Tāj Mahal. (C.M.N.)

SINHALESE LITERATURE: 10TH CENTURY AD TO 19TH CENTURY

The island nation of Ceylon (now called Sri Lanka), formally a part of South Asia, has been little noticed by the subcontinent, apart from the fact that according to an uncertain tradition it is celebrated in the *Rāmāyaṇa* as the island called Laṅkā. Buddhist sway was introduced there early, during the reign of Aśoka Maurya (*c.* 269–232 BC); and, while on the subcontinent Buddhism prospered, declined, and finally disappeared, in Ceylon it has continued until today. Although there are obvious borrowings in Ceylon from subcontinental literature, notably Sanskrit, and there was rather precarious communication with India through the island's Hindu community of Tamils, Ceylon never became culturally continuous with the mainland. The language itself, although of Indo-Aryan stock, is strongly mixed with a substratum of Dravidian. Also, it was Ceylon's fate early to fall victim to European colonialism, first to the Portuguese, then to the Dutch, and finally to the British, before it regained nationhood in 1948.

While there are inscriptions that antedate the Christian Era, no texts appear to survive from before the 10th century AD. The first texts that emerged were aids in Sinhalese—glossaries, paraphrases, and the like—to the study of the Pāli texts of Buddhism. More interesting are Sinhalese renderings of the life and virtues of the Buddha. Important in this genre, hagiographic rather than literary, is the *Amāvatura* ("Flood of the Ambrosia"), by Guruḷugōmī, which in 18 chapters purports to narrate the life of the Buddha, with specific emphasis on one of his nine virtues—his capacity to tame recalcitrant people or forces. In a similar vein is the literature of devotion and counsel, in which Buddhist virtues are celebrated.

Historical records

Exceptional in the context of the South Asian subcontinent is the early and persistent interest in historical records. Such interest had begun in Pāli with the *Dīpavaṃsa* ("Chronicle of the Island") and had continued with the *Mahāvaṃsa* ("Great Chronicle") and *Cūlavaṃsa* ("Lesser Chronicle"), but it had a life of its own in Sinhalese. The most important, and possibly the oldest, of such chronicles is the *Thūpavaṃsaya* ("Chronicle of the Great Stupa"), by Pārakrama Paṇḍita. Subsequent chronicles, or genealogies of places, comprise the history of all of the major Buddhist monuments. Several chronicles were also inspired by the Tooth Relic, received from Kaliṅga in the 4th century by King Kīrtiśrīmēghavarṇa. Such chronicling included that of the kings who protected the relic.

All of this literature was mostly in prose, but poetry as a literary form no doubt antedated it, as evidenced by early inscriptions. Much poetry was occasioned by Pāli *Jātakas* (stories of the Buddha's previous births) and other Buddhist stories, though Hindu stories were not lacking; for example, a version of the Sanskrit *Mahābhārata* (received through a Tamil source) was cast in the style of a *Jātaka* in the *Mahāpadaraṅga-Jatakaya.*

Likewise of Hindu Indian origin was a genre that took off from the Sanskrit poet Kālidāsa's "Meghadūta" (see above *Classical Sanskrit kāvya* [*200–1200*]), in which an exiled lover sends a message to his beloved by way of a monsoon cloud, thus giving the poet the opportunity to dwell on the description of landmarks in a poetic travelogue. This genre, so-called *saṃdeśa* literature, by no means unknown on the mainland, proliferated widely on Ceylon.

Of a different style are panegyrics and war poems, the earliest of which is the *Parakumbasirita* ("History of Parakramabahu VI," king in Jayavardhanapura from 1410 to 1468). Again reminiscent of the mainland and the religious tradition are the plentiful eulogies of the Buddha. Popular, too, was didactic verse, among the most notable of which is the *Kusajātaka,* 687 stanzas of epigrams and exempla by the 17th-century poet Alagiyavanna Mohoṭṭāla.

MODERN PERIOD: 19TH AND 20TH CENTURIES

The modern period was ushered in by the arrival of the British, the influence of Western models becoming discernible in the early 19th century. Reform-minded Hindus, led by Ram Mohun Roy, took a positive attitude to Western literature and urged on their countrymen a Western type of education. Newly formed literary clubs spread the influence of predominantly British works, thereby opening the Indian educated elite to Western culture and literature in general. After a period of translation, authors sought to imitate Western models and eventually to be independently creative in the new styles.

Effects of Westernization

The most striking result of Westernization was the introduction of prose on a major scale. Vernacular prose, rarely looked upon previously as a medium for art, was now used as a literary vehicle, and such hitherto unknown forms as the novel, novella, and short story began to emerge. In poetry the thrall of tradition was stronger, and verse in the older forms continued to be written. With modernity, realism appeared, as well as symbolism in some quarters, and there was new psychological and social interest.

From Bengal spread a new sense of national purpose, which became the principal motivation for much English as well as vernacular literature. Three trends can be distinguished in the products of this increasing literary activity. The old traditionalism was transformed into romanticism, which looked to the past, to Indian history, for inspiration and sought to preserve what was considered valuable in the past; a tendency to mysticism went hand in hand with the romantic mood (a mood that was also widespread in 19th-century Europe). Greater social awareness in European literature was reflected in the literature of Indian progressives, in whose works a somewhat romantic Marxism prevailed. Finally, there was a humanistic trend. The teachings of Mahatma Gandhi, combining social concerns with traditional ethics, later exerted a very great influence on literature.

In the years preceding and following India's independence (1947) and control of the princely states, the fervour of writers sometimes turned to an increasingly articulate progressivism of various Marxist schools, sometimes to disappointment and bitterness, and most recently, it appears, to a mood of introspection. These developments, which occurred with a different pace in different regions, are described briefly below. A complete coverage of the most modern literature has not been attempted, but an endeavour has been made to mention persons who are considered to be representative.

Bengali. Except for the iconoclastic poet Michael Madhusudan Datta, poetic activity in the mid-19th century was giving ground to experimentation with the new prose style learned from English. During this period, Bengali literature produced a spate of novels—satiric, social, and picaresque. While Michael's work *Mēghanādavadh* (1861; a long poem on the Rāma theme in which Rāma and Lakṣmaṇa become the villains and Rāvaṇa the hero) caused a stir, the literary event of the period was the appearance on the scene of Bankim Chandra Chatterjee, whose first novel, *Durgeśanandinī* ("Daughter of the Lord of the Fort"), appeared in 1865. While not at first overtly nationalist, Bankim Chandra became more and more an apologist for the Hindu position. In *Kṛṣṇacaritra,* Christ suffers in comparison with Krishna, and in his best known work, *Ānanda-maṭh* (1892; "The Abbey of Bliss"), the motherland in the person of the goddess Durgā is extolled.

Perhaps first among novelists of the late 19th and early 20th centuries is Saratchandra Chatterjee, whose social concerns with the family and other homely issues made his work popular. But the early 20th century is certainly best known for the poet who towers head and shoulders

above the rest, Rabindranath Tagore. Poet, playwright, novelist, painter, essayist, musician, social reformer, Rabindranath produced works, still not completely collected, that fill 26 substantial volumes. The winner of the Nobel Prize for Literature in 1913, primarily for his little book of songs called *Gītāñjali,* which was much praised by Ezra Pound and William Butler Yeats, Tagore is more known for these devotional poems than for the wit and clear thought with which his later work is filled. He was the last of an era, looking back as he did to the religious and political history of Bengal for his inspiration. Those who followed him were more concerned with introspection and dramatic imagery.

If Tagore was the last poet in the Bengali tradition, Jibanananda Das was the first of a new breed. Musing and melancholy, yet known for vivid and unusual imagery Jibananada is a poet who has much influence on younger writers in Bengal. There have been many other poets in the 20th century who are equally powerful but stand somewhat apart from the mainstream. One of these was Sudhindranath Datta, a poet much like Pound in careful and etymological use of language; another is the poet and prose writer Buddhadeva Bose.

Bose has been termed a progressive, and indeed he consciously turned away from the tradition orientation of Tagore and sought inspiration in schools foreign to Bengal—for example, the French Symbolists. He is the leader of an artistic faction, the Kallol school, and editor of an influential literary magazine, *Kavitā.* Unjustifiably called obscene, his writing has been experimental, probing into social and psychological realities of Bengali life.

While there have been, and still are, literary factions associated with political positions, they have been less definitive than some in other parts of India. Bengali writers in the 20th century have tended to be personal and individual rather than propagandist for political positions.

Assamese. Assamese literature began with Hemchandra Baruwa, a satirist and playwright, author of the play *Bahiri-Rang-Chang Bhitare Kowabhaturi* (1861; "All That Glitters Is Not Gold"). The most outstanding among the early modern writers was Lakshminath Bezbaruwa, who founded a literary monthly, *Jōnāki* ("Moonlight"), in 1889, and was responsible for infusing Assamese letters with 19th-century Romanticism. Later 20th-century writers have tried to remain faithful to the ideals of *Jōnāki.* The short story in particular has flourished in the language; notable practitioners are Mahichandra Bora and Holiram Deka.

The year 1940 marked a shift toward psychology, but World War II effectively put an end to literary development. When writers resumed after the war, there was a clear break with the past, in experimental verse and the growth of the novel form.

Hindi. Modern Hindi literature began with Harishchandra in poetry and drama, Mahavir Prasad Dvivedi in criticism and other prose writings, and Prem Chand in fiction. This period, the second half of the 19th century, saw mainly translations from Sanskrit, Bengali, and English. The growth of nationalism and social reform movements of the Arya Samaj led to the composition of long narrative poems, exemplified by those of Maithili Sharan Gupta; dramas, by those of Jayashankar Prasad; and historical novels, by those of Prasad, Chatureen Shastri, and Vrindavan Lal Varma. The novels drew mainly on the periods of the Maurya, Gupta, and Mughal empires.

Gandhi's influence on poetry

This period was followed by the Non-cooperation and *satyāgraha* movements of Mahatma Gandhi, which inspired poets such as Makhan Lal Chaturvedi, Gupta, and Subhadra Chauhan and novelists such as Prem Chand and Jainendra Kumar. Eventual disillusionment with Gandhian experiments and the increasing influence of Marxism on European literature influenced writers such as Yashpal, Rangeya Raghava, and Nagarjuna.

S.N. Pant, Prasad, Nirala, and Mahadevi Varma, the most creative poets of the 1930s, drew inspiration from the Romantic tradition in English and Bengali poetry and the mystic tradition of medieval Hindi poetry. Reacting against them were the Marxist poets Ram Vilas Sharma and Nagarjuna and experimentalists such as H.S. Vatsyayan "Agyeya" and Bharat Bhuti Agarwal. Nirala, who developed from a mystic-romantic into a realist and experimentalist, was the most outstanding poet of the 1950s; and Sarweshwar Dayal Saxena, Kailash Vajpeyi, and Raghubir Sahay were the most creative poets of the 1960s.

Two trends, represented by the work of Prem Chand and Jainendra Kumar, led Hindi fiction in two different directions: while social realists like Yashpal, Upendranath Ashk, Amritlal Nagar, Mohan Rakesh, Rajendra Yadav, Kamleshwar, Nagarjuna, and Renu faithfully analyzed the changing patterns of Indian society, writers such as Ila Chandra Joshi, "Agyeya," Dharm Vir Bharati, and Shrikant Varma explored the psychology of the individual, not necessarily within the Indian context.

Among the dramatists of the 1930s and 1940s were Govind Ballabh Pant and Seth Govind Das; because of their highly Sanskritized language, their plays have had a limited audience. Plays by minor writers such as Ramesh Mehta, however, are repeatedly staged by professional theatres. In between these extremes there are some notable playwrights.

Gujarati. In Gujarāt, too, the advent of British rule deeply influenced the literary scene. The year 1886 saw the *Kusumamālā* ("Garland of Flowers"), a collection of lyrics by Narsingh Rao. Other poets include Kalapi, Kant, and especially Nanalal, who experimented in free verse and was the first poet to eulogize Gandhi. Gandhi, himself a Gujarati, admonished poets to write for the masses and thus inaugurated a period of poetic concern with changes in the social order. Many incidents in Gandhi's life inspired the songs of poets. The Gandhi period in Gujarāt as elsewhere gave way to a period of progressivism in the class-conflict poetry of R.L. Meghani and Bhogilal Gandhi. In post-independence India, poetry has tended to become subjectivist and alienated without, however, fully superseding the traditional verse of devotion to God and love of nature.

Among novelists, Govardhanram stands out; his *Sarasvatīchandra* is a classic, the first social novel. In the novel form, too, the influence of Gandhiism is clearly felt, though not in the person of Kanaiyalal Munshi, who was critical of Gandhian ideology but still, in several *Purāṇa*-inspired works, tended to preach much the same message. In the period after independence the modernists embraced existentialistic, surrealistic, and symbolistic trends and gave voice to the same kind of alienation as the poets.

Marathi. The modern period in Marathi poetry began with Kesavasut and was influenced by 19th-century British Romanticism and liberalism, European nationalism, and the greatness of the history of Mahārāshtra. Kesavasut declared a revolt against traditional Marathi poetry and started a school, lasting until 1920, that emphasized home and nature, the glorious past, and pure lyricism. After that, the period was dominated by a group of poets called the Ravikiraṇ Maṇḍal, who proclaimed that poetry was not for the erudite and sensitive but was instead a part of everyday life. Contemporary poetry, after 1945, seeks to explore man and his life in all its variety; it is subjective and personal and tries to speak colloquially.

Among modern dramatists, S.K. Kolhatkar and R.G. Gadkari are notable. Realism was first brought to the stage in the 20th century, by Mama Varerkar, who tried to interpret many social problems.

Novelists of the 20th century

The *Madhalī Sthiti* (1885; "Middle State"), of Hari Narayan Apte, began the novel tradition in Marathi; the work's message was one of social reform. A high place is held by V.M. Joshi, who explored the education and evolution of a woman (*Susīlā-cha Diva,* 1930) and the relation between art and morals (*Indu Kāḷe va Saralā Bhoḷe,* 1935). Important after 1925 were N.S. Phadke, who advocated art for art's sake, and V.S. Khandehar, who countered the former with an idealistic art for life's sake. Noteworthy contemporary novelists are S.N. Pendse, V.V. Shirwadkar, G.N. Dandekar, and Ranjit Desai.

Punjabi. Modern Punjabi literature began around 1860. A number of trends in modern poetry can be discerned. To the more traditional genres of narrative poetry, mystic verse, and love poems was added nationalist poetry in a humorous or satiric mood and experimental verse. Among

the more important Punjabi poets are Bhai Vir Singh, in the 19th century, and Purana Singh, Amrita Pritam, and Baba Balwanta, in the 20th century.

Modern prose is represented by Bhai Vira Singha, Charana Singha, and Nanaka Singha, all of whom wrote novels; the same writers, as well as Gurbhaksh Singh and Devendra Satyarathi, also wrote short stories. Among playwrights mention may be made of I.C. Nanda, Harcharan Singh, and Santa Singh Sekhon.

Rajasthani. It is generally agreed that modern Rajasthani literature began with the works of Suryamal Misrama. His most important works are the *Vamsa Bhaskara* and the *Vira satsaī*. The *Vamsa Bhaskara* contains accounts of the Rājput princes who ruled in what was then Rājputāna (at present the state of Rājasthān), during the lifetime of the poet (1872–1952). The *Vira satsaī* is a collection of couplets dealing with historical heroes. Two other important poets in this traditional style are Bakhtavara Ji and Kaviraja Muraridana.

The period of nationalist strife against the British inspired a number of poets to verse that was both nationalist and in the traditional heroic vein; among them are Hiralala Sastri, Manikyalala Varma, and Jayanarayana Vyasa. This period was followed by one in which progressive social ideals inspired such poets as Ganeshilala Vyasa, Murlidhara Vyasa, and Satyaprakasha Jodhi.

Primarily known for their lyrics are Kanhaya Lal Sethiya and Megharaja Mukula, among others, and known for their narrative poems are Manohara Sharma, Shrimanta Kumara, and Naraina Singha Bhati.

Modern prose is represented in the novel, short story, and play. Among the novelists are Shiva Candra Bharatiya, Shri Lal Jodhi, Vijaya Dana Detha, and Yadavendra Sharma Chandra; the short-story writers are Rani Lakshmi Kumari Chandavata, Narasingh Rajapurohita, Dinadayala Ojha, and Purushottama Lala Menariya. Vijaya Dana Detha and Rani Lakshmi Kumari Chandavata are also known for their retelling of Rajasthani folktales. Among the playwrights is Shivachandra Bharatiya.

Tamil. In the second half of the 19th century two tendencies were present in Tamil literature. One was the old traditional prose style of the *Patiṉeṇ-kīḻkkaṇakku*, or "Eighteen Ethical Works" (see above *Dravidian literature: 1st–19th century*), learned and severely scholastic; among others, V.V. Svaminatha Iyer and Arumuga Navalar wrote in this style. Another tendency, begun by Aruṇācala Kavirāyar in the 18th century, sought to bring the spoken and written languages together. This tendency developed on one side into such works as the operatic play *Nantaṇār Carittarak Kīrttaṇai* by Gopalakrishna, and on the other into ballads, often based on the lore of the Sanskrit *Purāṇas*. Despite attempts to effect a synthesis between the two languages, however, the scholastic style has continued to have a profound influence on modern Tamil literature; the normal spoken language, in fact, never became a literary medium.

First Tamil novel

The first novel in Tamil appeared in 1879, the *Piratāpamutaliyār Carittiram*, by Vetanayakam Pillai, who was inspired by English and French novels. In important respects Pillai's work is typical of all early modern Tamil fiction: his subject matter is Tamil life as he observed it, the language is scholastic, and the inspiration comes from foreign sources. Not strictly a novel, his work, which has a predominantly moral tone, is a loosely gathered string of narratives centred around an innocent hero.

Quite different is the *Kamalāmpāḷ Carittiram* ("The Fatal Rumor"), by Rajam Aiyar, whom many judge to be the most important prose writer of 19th-century Tamil literature. In this work, the author created a series of characters that appear to have become classics; the story is a romance, yet life in rural Tamil country is treated very realistically, with humour, irony, and social satire. In language Aiyar follows the classical style, which he intermixes with informal conversation, a style that has been imitated by modern authors.

The turn of the century saw the development of the *centamiḻ* style, which in many respects is a continuation of the medieval commentatorial style. The best representative is V.V. Swaminathan, who also is responsible for the rediscovery of the Tamil classical legacy, usually called "Tamil Renaissance," which tended to direct the mood of writers back to the glorious past. The pride in Tamil subsequently gave rise to a purist tradition and a second style, called *tuyattamiḻ*, or "pure Tamil." With exaggerated Tamilian self-consciousness, the language was purged of all non-Tamil loanwords, particularly Sanskrit, which removed the literary language even further from the spoken one. This style was not ineffective in verse but led easily to rhetoric.

The purist trend brought forth a reaction in *putumaṇipravāla naṭai*, "the new *maṇipravāla*" (see above *Dravidian literature: 1st–19th century*), which was Sanskritized with a vengeance and is of little literary interest.

The scholastic and formalist character of Tamil prose was predominant in the literature until the advent, in the early 20th century, of the poet and prose writer Subrahmanya Bharati. Bharati sought to synthesize the popular and the scholastic traditions of Tamil literature, and he created thereby a Tamil that was amenable to all literary expression. This synthesis, however, did not extend to the literary language itself, which in grammar continued the formal language, though for syntax, vocabulary, etc., he drew upon colloquial speech. In doing so he saved the language from the Sanskrit tradition of *Purāṇa* writing. His style is the *maṟumaḻarcci naṭai*, the "renaissance style."

In the first half of the 20th century, R. Krishnamurthy was an immensely popular writer. Under the pseudonym Kalki, he was an influential journalist who wrote voluminous historical romances.

In the 1930s there was a literary movement inspired by a journal called *Manikkoti*. Writers in this movement contributed extremely important new works, both in verse and prose, to Tamil letters. Among them was Putumaippittan, who wrote realistically, critically, and even bitterly about the failings of society.

Contemporary literature is represented by T. Janakiraman, who writes novels, short stories, and plays with themes from urban Tamil middle-class family life; Jayakanthan, a sharp and passionate writer, with a tendency to shock his readers; and L.S. Ramatirthan, probably the finest stylist at work in Tamil today, who started by writing in English.

Malayalam. In Malayalam the modern movement began in the late 19th century with Asan, who was temperamentally a pessimist—a disposition reinforced by his metaphysics—yet all his life was active in promoting his downtrodden Ezhava community. Ullor wrote in the classical tradition, on the basis of which he appealed for universal love, while Vallathol (died 1958) responded to the human significance of social progress.

Contemporary poetry records the encounter with problems of social, political, and economic life. The tendency is toward political radicalism.

Drama, native in Malayalam tradition, emerged in the modern period as farce, comedy, and satire but turned in the 1920s to a more sombre appraisal of outdated social conventions. The novel dates back to the late 1880s and was early concerned with social realism. At present the general tendency is introspective.

Emergence of modern Kannada poetry

Kannada. Modern Kannada poetry emerged about the beginning of the 20th century and showed a spirit of national purpose that pervaded other literature as well. By 1920, after major translations from Western models had been published, new literary forms such as the lyric and the short story came to the fore in the works of Panje Mangesh Rao and B.M. Srikantiah. Other prominent Kannada writers were D.V.G. Masti, Govinda Pai, and K.V. Puttappa ("Kuvempu"). In recent years a modernist movement has influenced the literature.

Urdu. The modern period in Urdu literature coincides with the mid-19th-century emergence of a middle class that saw in Western thought and science a means to needed social reform. Naẕīr Aḥmad wrote novels about the conflicts of Muslim middle class people. Shiblī, a poet and critic, wrote on the lives of great Muslims. The more famous novelists of the later period are Ratan-Nāth Sharshār, ʿAbd-ul-Ḥalim Sharar, and Mīrzā Ruswā. The fathers of modern Urdu poetry were Ḥālī and Muḥammad

Ḥusayn Āzād, the latter particularly characterized by a fine sensitivity for the past.

The greatest modern poet is Iqbāl. Writing in the early 20th century, he was influenced by the general sense of national purpose and the freedom movement. His poetic imagery, the power of his expression, and his philosophical outlook won the admiration of his fellow Muslims. In prose the most important writer of short stories was Prem Chand, who late in his career took to writing in Hindi. The 1930s saw the influence of progressivism, which attempted to make literature an arm of social revolution. Among the representative writers of this period are Sajjad Zahir, Upendranath Ashk, and Ismat Chughtai, the last a woman who is considered among the best.

English. There has been Indian literary activity in English for the last 200 years. It began with the insistence of the reformist Rammohan Ray and other like-minded Hindus that, for India to take its rightful place among nations, a knowledge of and education in English were essential. English literary activity took on a new aspect with the independence movement, whose leaders and followers found in English the one language that united them.

Among the first poets were Henry Derozio, Kashiprasad Ghose, and Michael Madhusudan Datta, all of whom wrote narrative verse. In the following generation there was Toru Dutt, important among women poets in this genre. Carrying on her work was Sarojini Naidu, judged by many the greatest of women poets; among her writings are *The Golden Threshold* (1905), *The Bird of Time* (1912), and *The Broken Wing* (1917). Best known of the Indian poets in English was the Bengali Rabindranath Tagore (see above *Bengali*), who, however, wrote most of his verse first in Bengali, and then translated it. A very different figure from Tagore is Sri Aurobindo, who started out as an ardent nationalist and was jailed by the British. After his conversion from activism to introspection, which took place in jail, he established a hermitage in Pondicherry. He left behind a rich *oeuvre* of verse that has inspired a contemporary school of mystic poets. Other modern poets show the influence of T.S. Eliot and Ezra Pound.

The independence movement gave strong impetus to expository prose. Important contributors to this genre were Bal Gangadhar Tilak, who edited the English journal *Mahratta,* Lala Lajpat Rai, Kasturi Ranga Iyengar, and T. Prakasam. Mahatma Gandhi, too, wrote widely in English and edited *Young India* and the *Harijan.* He also wrote the autobiography *My Experiments with Truth* (originally published in Gujarati, 1927–29), now an Indian classic. In this he was followed by Jawaharlal Nehru, whose *Discovery of India* is justly popular.

Prose fiction

Prose fiction in English began in 1902 with the novel *The Lake of Palms,* by Romesh Chunder Dutt. The next important novelist is Mulk Raj Anand, who fulminated against class and caste distinction in a series of novels, *The Coolie* (1936), *Untouchable* (1935), *Two Leaves and a Bud* (1937), and *The Big Heart* (1945). Less fierce, though a better craftsman, is R.K. Narayan, who has published nine novels (as well as many short stories), among them *The Guide* (1958), *The Man-Eater of Malgudi* (1961), and *The Vendor of Sweets* (1967); his work has a wider circle of readers outside India than within. Other Indian novelists in the English medium are Santha Rama Rau, Manohar Malgonkar, Kamala Markandeya, and Khushwant Singh. The most popular is Raja Rao, whose novels *Kanthapura* (1938), *The Cow of the Barricades* (1947), and *The Serpent and the Rope* have attracted a wide following.

Sinhalese. Traditional contemporary poetry continues to be Buddhist in subject matter and sentiment. A more modern literature arose under the influence of Western models; notable among the contemporary representatives of Sinhalese literature are Kumaranatunga, a critic, Matin Wickremasinghe, a novelist, and Tennakoon, a poet.

(J.A.B.v.B.)

Music

FOLK, CLASSICAL, AND POPULAR MUSIC

Rural areas. The wide field of musical phenomena in South Asia ranges from the relatively simple two- or three-tone melodies of some of the hill tribes in central India to the highly refined art music heard in concert halls in the large cities. This variety reflects the heterogeneous population of the subcontinent in terms of race, religion, language, and social status. In the villages, music is not just a form of entertainment but is an essential element in many of the activities of daily life and plays a prominent part in most of the rituals. These include life-cycle events, such as birth, initiation, marriage, and death; events of the agricultural cycle, such as planting, transplanting, harvesting, and threshing; and a variety of work songs. Much of this music could be described as functional, for it serves a utilitarian purpose; for instance, a harvest song might well give thanks to God for a bountiful harvest, but underlying this is the idea that singing this song in its traditional manner will help to ensure that the next harvest will be equally fruitful. These songs are usually sung by all of the members participating in the activity and are not sung for a human audience. They are often sung in the form of leader and chorus, and the musical accompaniment, if any, is generally provided by drone instruments (those sustaining or reiterating a given note or notes), usually of the lute family, or percussion instruments, such as drums, clappers, and pairs of cymbals. Occasionally, a fiddle or flute might also accompany the singers, who often dance while they sing.

In each area and even within a single area, different social groups have their own individual songs whose origins are lost in antiquity. The songs are passed on from one generation to another, and in most cases the composers are unknown. Apart from folk songs, one also hears outdoor instrumental music in villages. The music is provided by an ensemble of varying size, which consists basically of an oboe type of instrument (usually a *sheh'nai* in North India and *nagaswaram* in the south) and a variety of drums. Sometimes straight, curved, or S-shaped horns may be added. These groups play at weddings, funerals, and religious processions. The musicians are professional or semiprofessional and usually belong to a very low caste. Such ensembles are found in tribal as well as folk societies and in villages as well as in cities.

Itinerant musicians

Other professional music is also found in the rural regions. Most areas are visited by religious mendicants, many of whom travel around the countryside singing devotional songs, accompanying themselves either with a one-, two-, or three-stringed lute that generally provides only a drone or with a frame (tamborine-like) drum. They carry with them a small begging bowl and maintain themselves entirely on what they receive in alms. There are also itinerant magicians, snake charmers, acrobats, and storytellers who travel in the rural areas, often providing the only entertainment available in the villages. Music is often involved in their acts, and the storyteller generally sings his tales, which may be taken from the two epics, the *Mahābhārata* and the *Rāmāyaṇa,* or from the *Purāṇas,* the legends that describe the adventures of the incarnations of God as they rid the world of evil. Sometimes the narrative songs are concerned with historical characters and describe the wars and the heroic deeds of the regional rulers. Some storytellers specialize in generally tragic stories of romance and of lovers.

During certain religious festivals, the villages might be visited by a travelling band of players who enact some of the mythological episodes connected with the festival. Such performances are accompanied by music and may also include dances. During the festivals villagers may visit neighbourhood shrines or temples, there encountering religious mendicants singing devotional songs and perhaps watching elaborate enactments of the episodes connected with the festival. Thus, the villagers become familiar with the mythological and philosophical aspects of their religion, in spite of the low level of literacy in the rural areas and the difficulties of communication, often limited to a narrow dirt road traversed by bullock carts.

In modern times, rural areas are being influenced to a greater extent by urban culture. The principal impact has come through the introduction of relatively inexpensive transistorized radios, which have found their way into fairly remote villages. In addition, travelling cinemas, set

up quickly and easily in tents, have visited the rural areas for some years. As a result, the traditional rural forms of music and dance are in the process of change.

Classical music. In the cities many different forms of music can be heard. Of these the best known in the West are the classical music of North India, including Pakistan, sometimes called Hindustani music, and that of South India, or Karnatic music. Both classical systems are supported by an extensive body of literature and elaborate musical theory. Until modern times, classical music was patronized by the princely courts and to some extent also by the wealthy noblemen. Since India gained independence in 1947, and with the abolition of the princely kingdoms, the emphasis has shifted to the milieu of large concert halls. The concertgoer, radio, and the cinema are now the main patrons of the classical musicians. In recent times the growth of university music programs, particularly involving classical music, has placed greater emphasis on music history and theory and has provided a further source of income for musicologists and musicians. The traditional system of private instruction, however, still continues to this day.

Raga and *tāla*

Classical music is based on two main elements, raga and *tāla*. The word raga is derived from a Sanskrit root meaning "to colour," the underlying idea being that certain melodic shapes, involving specific intervals of the scale, produce a continuity of emotional experience and "colour" the mind. Since neither the melodic shapes nor their sequence are fixed precisely, a raga serves as a basis for composition and improvisation. Indian music has neither modulation (change of key) nor changing harmonies; instead, the music is invariably accompanied by a drone that establishes the tonic, or ground note, of the raga and usually its fifth (*i.e.*, five notes above). These are chosen to suit the convenience of the main performer, as there is no concept of fixed pitch. While a raga is primarily a musical concept, specific ragas have acquired, particularly in North Indian music, a number of extramusical elements and are associated with particular periods of the day, seasons of the year, colours, deities, and specific moods.

The second element of Indian music, *tāla,* is best described as time measure and has two main constituents: the duration of the time measure in terms of time units that vary according to the tempo chosen; and the distribution of stress within the time measure. *Tāla,* like raga, serves as a basis for composition and improvisation.

Indian classical music is generally performed by small ensembles of not more than five or six musicians. Improvisation plays a major part in a performance, and great emphasis is placed on the creativity and sensitivity of the soloist. A performance of a raga usually goes through well-defined stages, beginning with an improvised melodic prelude that is followed by a composed piece set in a particular time measure. The composition is generally quite short and serves as a frame of reference to which the soloist returns at the conclusion of his improvisation. There is no set duration for the performance of a raga. A characteristic feature of North Indian classical music is the gradual acceleration of tempo, which leads to a final climax.

Nonclassical music of the cities. Classical music interests only a small proportion of the peoples of South Asia, even in the cities. Since about the 1930s a new genre, associated with the cinema, has achieved extraordinary popularity. Most Indian films are very much like Western musicals and generally include six or more songs. Film music derives its inspiration from a number of sources, both Indian and Western; classical, folk, and devotional music are the main Indian sources, while Western influence is seen most obviously in the use of large orchestras that employ both Western and Indian instruments. The influence of Western popular music, too, is very evident. In spite of the eclectic nature of Indian film music, most of the songs maintain an Indian feeling that arises largely from the vocal technique of the singers and the ornamentation of the melody line. This music is an experimental and developing form, and there have been attempts to add harmony and counterpoint, some of which may seem rather naïve to the Western ear. But the film music differs from typical Western music in that the melody line is generally not dictated by harmonic progressions and in that the harmonies used are incidental additions.

Film music

Aside from classical and film music, there are several other forms of urban music, some of which closely resemble the music of the rural areas. In city streets one is likely to encounter an outdoor band of oboes and drums announcing a wedding or a funeral. Street musicians, religious mendicants, snake charmers, storytellers, and magicians perform at every available opportunity, and work songs are sung by construction workers and other labourers. In private homes still other forms of music are performed, ranging from religious chanting to traditional folk and devotional songs. In public places of entertainment, the listener may encounter, apart from classical and film music, theatrical music from one of the relatively modern forms of regional theatre; and in the lowbrow places of entertainment courtesans still sing and dance in traditional fashion. In the larger cities there are performances of Western chamber music and occasionally symphony concerts, as well as popular dance music, rock, and jazz in the night clubs.

ANTIQUITY

In a musical tradition in which improvisation predominates, and written notation, when used, is skeletal and more a tool of the theorist than of the practicing musician, the music of past generations is irrevocably lost. References to music in ancient texts, aesthetic formulations, and depictions and written discussions of musical instruments can offer clues. In rare instances an ancient musical style may be preserved in unbroken oral tradition. For most historical eras and styles, surviving treatises explaining musical scales and modes—the framework of melody—provide a particularly important means of recapturing at least a suggestion of the music of former times, and tracing the musical theory of the past makes clear the position of the present musical system.

Little is known of the musical culture of the Indus Valley civilization of the 3rd and 2nd millennia BC. Some musical instruments, such as the arched, or bow-shaped, harp and more than one variety of drum, have been identified from the small terra-cotta figures and among the pictographs on the seals that were probably used by merchants. Further, it has been suggested that a bronze statuette of a dancing girl represents a class of temple dancers similar to those found much later in Hindu culture. It is known that the Indus civilization had established trade connections with the Mesopotamian civilizations, so that it is possible that the bow harp found in Sumeria would also have been known in the Indus Valley.

Vedic chant. *Compilation of hymns.* It is generally thought among scholars that the Indus Valley civilization was terminated by the arrival of bands of semi-nomadic tribesmen, the Aryans, who descended into India from the northwest, probably in the first half of the 2nd millennium BC. An important aspect of Aryan religious life was the bard-priest who composed hymns in praise of gods, to be sung or chanted at sacrifices. This tradition was continued in the invaders' new home in northern India until a sizable body of oral religious poetry had been composed. By about 1000 BC this body of chanted poetry had apparently grown to unmanageable proportions, and the best of the poems were formed into an anthology called Rigveda, which was then canonized. It was not committed to writing, but text and chanting formula were carefully handed down by word of mouth from one generation to the next, up to the present period. The poems in the Rigveda are arranged according to the priestly families who used and, presumably, had composed the hymns. Shortly after this a new Veda, called the Yajurveda, basically a methodical rearrangement of the verses of the Rigveda with certain additions in prose, was created to serve as a kind of manual for the priest officiating at the sacrifices. At approximately the same time, a third Veda, the Sāmaveda, was created for liturgical purposes. The Sāmaveda was also derived from the hymns of the Rigveda, but the words were distorted by the repetition of syllables, pauses, prolongations, and phonetic changes.

as well as the insertion of certain meaningless syllables believed to have magical significance. A fourth Veda, the Atharvaveda, was accepted as a Veda considerably later and is quite unrelated to the other three. It represents the more popular aspects of the Aryan religion and consists mostly of magic spells and incantations.

Each of these Vedas has several ancillary texts, called the *Brāhmaṇas, Āraṇyakas,* and *Upaniṣads,* which are also regarded as part of the Vedas. These ancillary texts are concerned primarily with mystical speculations, symbolism, and the cosmological significance of the sacrifice. The Vedic literature was oral and not written down until very much later, the first reference to a written Vedic text being in the 10th century AD. In order to ensure the purity of the Vedas, the slightest change was forbidden, and the priests devised systems of checks and counterchecks, so that there has been virtually no change in these texts for about 3,000 years. Underlying this was the belief that the correct recitation of the Vedas was "the pivot of the universe" and that the slightest mistake would have disastrous cosmic consequence unless expiated by sacrifice and prayer. The Vedas are still chanted by the Brahmin priests at weddings, initiations, funerals, and the like, in the daily devotions of the priests, and at the now rarely held so-called public sacrifices.

Ritual accuracy of recitation

From the Vedic literature it is apparent that music played an important part in the lives of the Aryan peoples, and there are references to stringed instruments, wind instruments, and several types of drums and cymbals. Songs, instrumental music, and dance are mentioned as being an integral part of some of the sacrificial ceremonies. The bow harp (*vīṇā*), a stringed instrument (probably a board zither) with 100 strings, and the bamboo flute were the most prominent melody instruments. Little is known of the music, however, apart from the Vedic chanting, which can still be heard today.

Chant intonation. The chanting of the Rigveda and Yajurveda shows, with some exceptions, a direct correlation with the grammar of the Vedic language. As in ancient Greek, the original Vedic language was accented, with the location of the accent often having a bearing on the meaning of the word. In the development of the Vedic language to Classical Sanskrit, the original accent was replaced by an automatic stress accent, whose location was determined by the length of the word and had no bearing on its meaning. It was thus imperative that the location of the original accent be inviolate if the Vedic texts were to be preserved accurately. The original Vedic accent occurs as a three-syllable pattern: the central syllable, called *udātta,* receives the main accent; the preceding syllable, *anudātta,* is a kind of preparation for the accent; and the following syllable, *svarita,* is a kind of return from accentuation to accentlessness. There is some difference of opinion among scholars as to the nature of the original Vedic accent; some have suggested that it was based on pitch, others on stress; and one theory proposes that it referred to the relative height of the tongue.

In the most common style of Rigvedic and Yajurvedic chanting found today, that of the Tamil Aiyar Brahmins, it is clear that the accent is differentiated in terms of pitch. This chanting is based on three tones; the *udātta* and the nonaccented syllables (called *pracaya*) are recited at a middle tone, the preceding *anudātta* syllable at a low tone, and the following *svarita* syllable either at the high tone (when the syllable is short) or as a combination of middle tone and high tone. The intonation of these tones is not precise, but the lower interval is very often about a whole tone, while the upper interval tends to be slightly smaller than a whole tone but slightly larger than a semitone. In this style of chanting the duration of the tones is also relative to the length of the syllables, the short syllables generally being half the duration of the long.

Verbal accent and musical pitch

The more musical chanting of the Sāmaveda employs five, six, or seven tones and is said to be the source of the later secular and classical music. From some of the phonetic texts that follow the Vedic literature, it is apparent that certain elements of musical theory were known in Vedic circles, and there are references to three octave registers (*sthāna*), each containing seven notes (*yama*). An auxiliary text of the Sāmaveda, the *Nāradīśikṣā,* correlates the Vedic tones with the accents described above, suggesting that the Samavedic tones possibly derived from the accents. The Samavedic hymns as chanted by the Tamil Aiyar Brahmins are based on a mode similar to the D mode (D-d on the white notes of the piano; *i.e.,* the ecclesiastical Dorian mode). But the hymns seem to use three different-sized intervals, in contrast to the two sizes found in the Western church modes. They are approximately a whole tone, a semitone, and an intermediate tone. Once again, the intervals are not consistent and vary both from one chanter to another and within the framework of a single chant. The chants are entirely unaccompanied by instruments, and this may account for some of the extreme variation of intonation.

The changes brought by the 20th century have weakened the traditional prominent position of the Vedic chant. The Atharvaveda is seldom heard in India now. Samavedic chant, associated primarily with the large public sacrifices, also appears to be dying out. Even the Rigveda and Yajurveda are virtually extinct in some places, and South India is now the main stronghold of Vedic chant.

The classical period. The ritual of the Vedas involves only the three upper classes, or castes, of Aryan society: the Brahmin, or priestly class; the Kṣatriya, or prince-warriors; and the Vaiśya, or merchants. The fourth caste, the Śūdras, or labourers, were excluded from Vedic rites. The primary sources of religious education and inspiration for the Śūdra were derived from what is sometimes called the fifth Veda: the epic poems *Rāmāyaṇa* and *Mahābhārata,* as well as the collections of legends, called the *Purāṇas,* depicting the lives of the various incarnations of the Hindu deities. The *Rāmāyaṇa* and the *Mahābhārata* were originally secular in character, describing the heroic deeds of kings and noblemen, many of whom are not recorded in history. Subsequently, religious matter was added, including the very famous sermon *Bhagavadgītā* ("Song of the Lord"), which has been referred to as the most important document of Hinduism; and many of the heroes of the epics were identified as incarnations of the Hindu deities. The legends were probably sung and recited by wandering minstrels and bards even before the advent of the Christian Era, in much the same way as they still are. The stories were also enacted on the stage, particularly at the time of the religious festivals. The earliest extant account of drama is to be found in the *Nāṭya-śāstra* ("Treatise on the Dramatic Arts"), a text that has been dated variously from the 2nd century BC to the 5th century AD and even later. It is virtually a handbook for the producer of stage plays and deals with all aspects of drama, including dance and music.

Music and drama

Theatrical music of the period apparently included songs sung on stage by the actors, as well as background music provided by an orchestra (which included singers) located offstage, in what was very like an orchestra pit. Melodies were composed on a system of modes, or *jātis,* each of which was thought to evoke one or more particular sentiments (*rasa*) by its emphasis on specific notes. The modes were derived in turn from the 14 *mūrchanās*—seven pairs of ascending seven-note series beginning on each of the notes of two closely related heptatonic (seven-note) parent scales, called *ṣaḍjagrāma* and *madhyamagrāma.* The *mūrchanās* were thus more or less analogous to the European modal scales that begin progressively on D, E, F, G, etc. A third parent scale, *gāndhāragrāma,* was mentioned in several texts of the period and some even earlier but is not included in the system laid out in the *Nāṭya-śāstra.*

Qualities of the scales. The two parent scales differed in the positioning of just one note, which was microtonally flatter in one of the scales. The microtonal difference, referred to as *pramāṇa* ("measuring") *śruti,* presumably served as a standard of measurement. In terms of this standard it was determined that the intervals of the *mūrchanās* were of three different sizes, consisting of two, three, or four *śrutis,* and that the octave comprised 22 *śrutis.* An interval of one *śruti* was not used. Several modern scholars have suggested that the *śrutis* were of unequal size; from the evidence in the *Nāṭya-śāstra,* it would appear, however, that they were thought to be equal. There

has been no attempt to determine the exact size of the *śrutis* in any of the traditional Indian musical treatises until relatively modern times (18th century). The term *śruti* was also used to define consonance and dissonance, as these terms were understood in the period. In this connection, four terms are mentioned: *vādī,* comparable to the Western term sonant, meaning "having sound"; *saṃvādī,* to the Western consonant (concordant; reposeful); *vivādī,* to dissonant (discordant; lacking repose); and *anuvādī,* to assonant (neither consonant nor dissonant). As in the ancient Greek Pythagorean system, which influenced Western music, only fourths and fifths (intervals of four or five tones in a Western scale) were considered consonant. In the Indian system of measurement, tones separated by either nine or 13 *śrutis* correspond in size to Western fourths and fifths and are described as being consonant to each other. "Dissonant" in this system referred only to the minor second, an interval of two *śrutis,* and to its inversion (complementary interval), the major seventh (20 *śrutis*). All other tones, including the major third, were thought to be assonant.

Melodically important microtonal differences

The musical difference between the two parent scales is best seen not in terms of the microtonal deviation mentioned earlier but rather in terms of a musically influential consonance found in one but lacking in the other and vice versa. In each of the parent scales there are two nonconsonances, one of which is the tritone (interval of three Western whole tones, such as F-B) of 11 *śrutis* inevitable in all diatonic scales (seven-note scales of the major scale and *mūrchanā* type) and which in Europe during the Middle Ages was described as *diabolus in musica* ("the devil in music").

The second is a microtonal nonconsonance unique to this ancient Indian system. It can be illustrated by referring in the subsequent explanation to Table 1, in which the seven Indian notes *ṣaḍja, ṛṣabha, gāndhāra, madhyama, pañcama, dhaivata,* and *niṣāda* are given in their commonly abbreviated forms, *ṣa, ṛi, ga, ma, pa, dha,* and *ni.*

The nonconsonance arises from variances of one *śruti* from the fundamental consonances of the fourth and the fifth—a variance of about a quarter tone. In the *ṣaḍjagrāma* scale the interval *ṛi-pa* (E⁻ to A) contains 10 *śrutis; i.e.,* one more than the nine of the consonant fourth. Comparably, in the *madhyamagrāma* scale the interval *ṣa-pa* (D to A⁻) contains 12 *śrutis,* or one fewer than the consonant fifth. These variances involve the consonant relationships of two melodically prominent notes, the first and the fifth. In the *madhyamagrāma* the first note, *ṣa,* has no consonant fifth, and perhaps for this reason this scale is said to begin not on the *ṣa* (D) but on its fourth, the note *ma* (G); hence, it resembles the G mode—*i.e.,* the ecclesiastical Mixolydian mode—whereas the *ṣaḍjagrāma* resembles the D mode, the ecclesiastical Dorian.

There is a striking resemblance of the *ṣaḍjagrāma* scale to the intervals used by the Tamil Aiyar Brahmins in their chanting of the Sāmaveda. Not only are their hymns set in a mode similar to the D mode, but they seem to use three different-sized intervals, the intermediate one corresponding to the three-*śruti* interval. The *Nāṭya-śāstra* claims to have derived song (*gīta*) from the chanting of the Sāmaveda, and the resemblances between the two may not be entirely fortuitous.

The two parent scales are complementary and between them supply all the consonances found in the ancient Greek Pythagorean scale. Thus, if in a mode the consonance *ṛi-pa* (E–A) were needed, one would tune to the *madhyamagrāma* scale. But, if the consonance *ṣa-pa* (D–A) were important, it could be obtained with the *ṣaḍjagrāma* tuning. There was a further development in this system caused by the introduction of two additional notes, called *antara ga* (F♯) and *kākalī ni* (C♯), which could be substituted for the usual *ga* (F) and *ni* (C). The *antara ga* eliminates the 11-*śruti* tritone between *ga* and *dha* (F–B), but its use creates a further tritone between F♯ and C. The second additional note, *kākalī ni* (C♯), eliminates this tritone but once again creates a new one, this time between C♯ and G. This process of adding notes, if carried further, would eventually lead to the circle, or, rather, the spiral, of fourths or fifths found in Western music (whereby a sequence of fifths, such as C–G, G–D, D–A, etc., leads eventually back to a microtonally out-of-tune C); there is no evidence that such a circle or spiral was known in ancient India.

Introduction of new notes

Mode, or jāti. From each of the two parent scales were derived seven modal sequences (the *mūrchanās* described above), based on each of the seven notes. The two *mūrchanās* of a corresponding pair differered from each other only in the tuning of the note *pa* (A), the crucial distinction in the tunings of the two parent scales. One of each pair was selected as the basis for a "pure" mode, or *śuddha-jāti;* in the groups of seven pure modes, four used the tuning of the *ṣaḍjagrāma* and three that of the *madhyamagrāma.* In addition to these seven pure modes, a further 11 "mixed" modes, or *vikṛta-jātis,* are also mentioned in the *Nāṭya-śāstra.* These were derived by a combination of two or more pure modes, but the text does not explain just in what way these derivations were accomplished.

The *jātis* were similar to the modern concept of raga in that they provided the melodic basis for composition and, presumably, improvisation. They were not merely scales, but were also assigned 10 melodic characteristics: *graha,* the initial note; *aṃśa,* the predominant note; *tāra,* the note that forms the upper limit; *mandra,* the note that forms the lower limit; *nyāsa,* the final note; *apanyāsa,* the secondary final note; *alpatva,* the notes to be used infrequently; *bahutva,* the notes to be used frequently; *ṣāḍavita,* the note that must be omitted in order to create the hexatonic (six-note) version of the mode; and *auḍavita,* the two notes that must be omitted to create the pentatonic (five-note) version of the mode.

No written music survives from this early period. It is not clear from the description whether or not the music was like that of the present period. There is no mention of a drone, nor do the instruments of the orchestra—consisting of the *vipañcī* and *vīṇā* (bow harps?), bamboo flute, a variety of drums, and singers—appear to include any specifically drone instrument, such as the modern tamboura. The evidence tends rather to suggest, from the emphasis on consonance and some of the playing techniques, that some form of organum (two or more parts paralleling the same melody at distinct pitch levels) and even some type of rudimentary harmony may have been characteristic.

Table 1: Intervals of Ṣaḍjagrāma and Madhyamagrāma Parent Scales*

Ṣaḍjagrāma

				10 *śrutis*							
Indian notes:	(*ni*)		*ṣa*	*ṛi*	*ga*	*ma*	*pa*	*dha*	*ni*		(*ṣa*)
Śruti intervals:		4	3	2	4	4	3	2	(4)		
Comparable Western notes†:	(C)		D	E⁻	F	G	A	B⁻	c		(d)
					11 *śrutis*						

Madhyamagrāma

				12 *śrutis*							
Indian notes:	(*ni*)		*ṣa*	*ṛi*	*ga*	*ma*	*pa*	*dha*	*ni*		(*ṣa*)
Śruti intervals:		4	3	2	4	3	4	2	(4)		
Comparable Western notes†:	(C)		D	E	F	G	A⁻	B⁻	c		(d)
					11 *śrutis*						

*Minus signs indicate slightly lower pitch. †Without reference to precise pitch.

MEDIEVAL PERIOD

Precursors of the medieval system. It is not clear just when the *jāti* system fell into disuse, for later writers refer to *jāti*s merely out of reverence for Bharata, the author of the *Nāṭya-śāstra.* Later developments are based on musical entities called *grāma-rāga*s, of which seven are mentioned in the 7th-century Kuṭimiyāmalai rock inscription in Tamil Nadu state. Although the word *grāma-rāga* does not occur in the *Nāṭya-śāstra,* the names applied to the individual *grāma-rāga*s are all mentioned. Two of them, *ṣaḍjagrāma-rāga* and *madhyamagrāma-rāga,* are obviously related to the parent scales of the *jāti* system. The other five seem to be variants of these two *grāma-rāga*s in which either or both the altered forms of the notes *ga* and *ni* (F♯ and C♯) are used. In the *Nāṭya-śāstra,* the reference to the various *grāma-rāga*s is far removed from the main section in which the *jāti* system is discussed, and there is no obvious connection between the two. Each of the *grāma-rāga*s is said to be used in one of the seven formal stages of Sanskrit drama. They have been reconstructed as shown in Table 2.

Character and possible origin of the raga

Further development of the grāma-rāgas. In the next significant text on Indian music, the *Bṛhaddeśī,* written by the theorist Mātaṅga about the 10th century AD, the *grāma-rāga*s are said to derive from the *jāti*s. In some respects, at least, the *grāma-rāga*s resemble not the *jāti*s but their parent scales. The author of the *Bṛhaddeśī* claims to be the first to discuss the term raga in any detail. It is clear that raga was only one of several kinds of musical entities in this period and is described as having "varied and graceful ornaments, with emphasis on clear, even, and deep tones and having a charming elegance." The ragas of this period seem to have been named after the different peoples living in the various parts of the country, suggesting that their origin might lie in folk music. Mātaṅga appears to contrast the two terms *mārga* and *deśī. Mārga* (literally "the path") apparently refers to the ancient traditional musical material, whereas *deśī* (literally "the vulgar dialect spoken in the provinces") designates the musical practice that was evolving in the provinces, which may have had a more secular basis. Although the title *Bṛhaddeśī* ("The Great Deśī") suggests that the latter music might have been the focus of the treatise and that the *grāma-rāga*s were possibly out of date by the time it was written, the surviving portion of the text does not support such a theory.

The mammoth 13th-century text *Saṅgītaratnākara* ("Ocean of Music and Dance"), composed by the theorist Śārṅgadeva, is often said to be one of the most important landmarks in Indian music history. It was composed in the Deccan (south central India) shortly before the conquest of this region by the Muslim invaders and thus gives an account of Indian music before the full impact of Muslim influence. A large part of this work is devoted to *mārga*—that is, the ancient music that includes the system of *jāti*s and *grāma-rāga*s—but Śārṅgadeva mentions a total of 264 ragas. Despite the use in both the *Bṛhaddeśī* and the *Saṅgītaratnākara* of a notation equivalent to the Western tonic sol–fa (*i.e.,* with syllables, as do–re–mi . . .) to illustrate the ragas, modern scholars have not yet been able to reconstruct them with assurance.

The basic difficulty scholars face lies in determining the intervals used in each of the ragas. In the ancient system, the *jāti*s were something like the ancient Greek and medieval church modes in that each was derived from a parent scale by altering the ground note and the tessitura (range). In modern Indian music, however, the ragas are all transposed to a common ground note. This change may well be connected with the introduction of the drone and the evolution of the long-necked-lute family on which the drone is usually played. In the old system, with the changing ground note, it would have been necessary to retune drone instruments from one raga to another, which would have been a cumbersome and impractical operation to carry out during a recital. It may have been this factor that provided the impetus for the change to the standard-ground-note system. There is no conclusive evidence to show just when this change might have taken place, and it is not clear whether the *Bṛhaddeśī* and the *Saṅgītaratnākara* are using the old ground-note system or one similar to that used in modern times.

THE ISLĀMIC PERIOD

Impact on musical genres and aesthetics. The Muslim conquest of India can be said to begin in the 12th century, although Sind (now in Pakistan) had been conquered by the Arabs as early as the 8th century. Muslim writers such as al-Jāḥiẓ and al-Masʿūdī had already commented favourably on Indian music in the 9th and 10th centuries, and the Muslims in India seem to have been very much attracted by it.

In the beginning of the 14th century the great poet Amīr Khosrow, who was considered to be extremely proficient in both Persian and Indian music, wrote that Indian music was superior to the music of any other country. Further, it is stated that, after the Muslim conquest of the Deccan under Malik Kāfūr (*c.* 1310), a large number of Hindu musicians were taken with the royal armies and settled

Table 2: Grāma-Rāgas

	scale												
Madhyama-grāma-rāga													
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		2		4		3	4		2		4	
Comparable Western notes	D	E⁻		F		G	A⁻		B⁻		c		d
Ṣaḍja-grāma-rāga													
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		2		4		4	3		2		4	
Comparable Western notes	D	E⁻		F		G	A		B⁻		c		d
Ṣādava-grāma-rāga				(a)*									
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		4		2		3	4		2		4	
Comparable Western notes	D	E⁻		F♯		G	A⁻		B⁻		c		d
Pañcama-grāma-rāga				(a)*									
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		4		2		4	3		2		4	
Comparable Western notes	D	E⁻		F♯		G	A		B⁻		c		d
Kaiśika-grāma-rāga				(a)*							(k)*		
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		4		2		3	4		4		2	
Comparable Western notes	D	E⁻		F♯		G	A⁻		B⁻		c♯		d
Sādhārita-grāma-rāga				(a)*							(k)*		
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*		*ṣa*
Śruti values	3		4		2		4	3		4		2	
Comparable Western notes	D	E⁻		F♯		G	A		B⁻		c♯		d
Kaiśika-madhyama-grāma-rāga				(a)*							(k)*		
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*			*dha*		*ni*		*ṣa*
Śruti values	3		4		2		7			4		2	
Comparable Western notes	D	E⁻		F♯		G			B⁻		c♯		d

*(a) and (k) refer to *antara* and *kākalī,* the variant forms of the notes *ga* and *ni.*

in the North. Although orthodox Islām considered music illegal, the acceptance of the Ṣūfī doctrines, in which music was an accepted means to the realization of God, enabled Muslim rulers and noblemen to extend their patronage to this art. At the courts of the Mughal emperors Akbar, Jahāngīr, and Shāh Jahān, music flourished on a grand scale. Apart from Indian musicians, there were also musicians from Persia, Afghanistan, and Kashmir in the employ of these rulers; nevertheless, it appears that it was Indian music that was most favoured. Famous Indian musicians, such as Svāmī Haridās and Tānsēn, are legendary performers and innovators of this period. After the example set by Amīr Khosrow, Muslim musicians took an active interest in the performance of Indian music and added to the repertoire by inventing new ragas, *tāla*s, and musical forms, as well as new instruments.

The Muslim patronage of music was largely effective in the north of India and has had a profound influence on North Indian music. Perhaps the main result of this influence was to de-emphasize the importance of the words of the songs, which were mostly based on Hindu devotional themes. In addition, the songs had been generally composed in Sanskrit, a language that had ceased to be a medium of communication except among scholars and priests. Sanskrit songs were gradually replaced by compositions in the various dialects of Hindi, Braj Bhasa, Bhojpuri, and Dakhani, as well as in Urdu and Persian; nevertheless, the problems of communication, in terms of both language and subject matter, were not easily reconciled. A new approach to religion was, in any case, sweeping through India at about this time. This emphasized devotion (*bhakti*) as a primary means to achieving union with God, bypassing the traditional Hindu beliefs of the transmigration of the soul from body to body in the lengthy process of purification before it could achieve the Godhead. The Islāmic Ṣūfī movement was based on an approach similar to that of the *bhakti* movements and also gained many converts in India. A manifestation of these devotional cults was the growth of a new form of mystic-devotional poetry composed by wandering mendicants who had dedicated their lives to the realization of God. Many of these mendicants have been sanctified and are referred to as poet-saints or singer-saints, since their poems were invariably set to music. A number of devotional sects sprang up all over the country, some Muslim, some Hindu, and others merging elements from both. These sects emphasized the individual's personal relationship with God. In their poetry, man's love for God was often represented as a woman's love for man and, specifically, the love of the milkmaid Rādhā for Krishna, a popular incarnation of the Hindu god Vishnu. In the environment of the royal courts, there was a less idealistic interpretation of the word love, and much of the poetry, as well as the miniature painting, of the period depicts the states of experience of the lover and the beloved.

Influence of religious movements

This attitude is also reflected in the musical literature of the period. From early times, both *jāti*s and ragas in their connection with dramatic performance were described as evoking specific sentiments (*rasa*) and being suitable for accompanying particular dramatic events. It was this connotational aspect, rather than the technical one, that gained precedence in this period. The most popular method of classification was in terms of ragas (masculine) and their wives, called *rāgiṇī*s, which was extended to include *putra*s, their sons, and *bhāryā*s, the wives of the sons. The ragas were personified and associated with particular scenes, some of which were taken from Hindu mythology, while others represented aspects of the relationship between two lovers. The climax of this personification is found in the *rāgamāla* paintings, usually in a series of 36, which depict the ragas and *rāgiṇī*s in their emotive settings.

Theoretical developments. From the middle of the 16th century, a new method of describing ragas is found in musical literature. It was also at about this time that the distinction between North and South Indian music became clearly evident. In the literature, ragas are described in terms of scales having a common ground note. These scales were called *mela* in the South and *mela* or *thāṭa* in the North.

Emergence of modern theoretical concepts

It was in the South that a complete theoretical system of *mela*s was introduced, in the *Caturdaṇḍiprakāśika* ("The Illuminator of the Four Pillars of Music"), a text written in the middle of the 17th century. This system was based on the permutations of the tones and semitones, which had by this time been reduced to a basic 12 in the octave. The octave was divided into two tetrachords, or four-note sequences, C–F and G–c, and six possible tetrachord species were arranged in an order showing their relationship with each other. It will be noted in the sequence of tetrachords shown below that each lower tetrachord has an analogous upper tetrachord and that the outer notes of each are constant, whereas the inner notes change their pitch.

1.	C	D♭	E♭♭	F	and	G	A♭	B♭♭	c
2.	C	D♭	E♭	F		G	A♭	B♭	c
3.	C	D♭	E	F		G	A♭	B	c
4.	C	D	E♭	F		G	A	B♭	c
5.	C	D	E	F		G	A	B	c
6.	C	D♯	E	F		G	A♯	B	c

The list could have extended further, except that apparently no pitch distinction was made between the enharmonic pairs D–E♭♭, D♯–E♭, A–B♭♭, and A♯–B♭. (Enharmonic notes have different pitch names but sound either the same pitch or, in some tuning systems, have very slight differences in pitch.)

By utilizing all possible combinations of a lower with an upper tetrachord, 36 *mela*s, or raga scales, were derived; a further 36 were formed by using F♯ in place of the F in the lower tetrachord. The *mela*s were named in such a way that the first two syllables of the name, when applied in a code, gave the number of that *mela* in the sequence. The musician, given the number, could easily reconstruct the scale of the *mela*. The names of the *mela*s were often derived from prominent ragas in those *mela*s, with a two-syllable prefix that supplied the code numbers; for instance, the name of the *mela Dhīra-śaṅkarābharaṇa* is derived from the raga *Śaṅkarābharaṇa*, the two syllables *dhīra* giving the code number 29, which indicates a scale similar to the Western major scale, or C mode. The *Caturdaṇḍiprakāśika* acknowledges the theoretical nature of its analytical system and mentions clearly that only 19 of the possible 72 *mela*s were in use at the time that the text was written.

Although North Indian texts also describe ragas in terms of *mela*s or *thāṭa*s, there is no attempt to arrange them systematically. In the *Rāgataraṅgiṇi* ("The River of Rāga"), probably of the 16th century, 12 *mela*s are mentioned:

bhairavī	C	D	E♭	F	G	A	B♭	c
toṛī	C	D♭	E♭	F	G	A♭	B♭	c
gaurī	C	D♭	E	F	G	A♭	B	c
karṇāta	C	D	E	F	G	A	B♭	c
kedāra	C	D	E	F	G	A	B	c
imana	C	D	E	F♯	G	A	B	c
sāraṅga	C	D	E♯	F♯	G	A♯	B	c
megha	C	D	E	F	G	A♯	B	c
dhanāśrī	C	D♭	E	F♯	G	A♭	B	c
pūravā	C	D	E	F♯	G	A⁺	B	c
mukhārī	C	D	E♭	F	G	A♭	B♭	c
dīpaka	no description							

Although it appears from the description of *sāraṅga* and *megha mela*s that enharmonic intervals were used, there is good reason to believe that the E♯ and A♯ in the two *mela*s really represent their chromatic counterparts, F and B♭, and that F and F♯ (and B and B♭) do not appear in sequence. The A⁺ in the *mela pūravā* is said to be raised by one *śruti*. The description of the ragas in these *mela*s shows that the North Indian system was by this time also based on 12 semitones.

THE MODERN PERIOD

With the collapse of the Mughal Empire in the 18th century and the emergence of the British as a dominant power in India, the subcontinent was divided into many princely states. Music continued to be patronized by the rulers, although the courts were never again to achieve their former opulence.

Musically, there has been a continuous evolution from the Islāmic period to the present, and both North and

Influence of theory on composition

South Indian classical music have continued to expand. South Indian music has clearly been influenced more by theory than has the North. The 72-*mela* system continues to be the basis of classifying the ragas in South India, but it has had more than a classificatory significance. Many new ragas have been composed in the past few centuries, some of them inspired by the theoretical scales of the *mela* system. As a result, there are now ragas in all of the 72 *mela*s.

In North Indian music, theory has had little influence on performance practice. This can be ascribed to the language problem, an especially significant influence on the many Muslim musicians in North India, who were not able to cope with the Sanskrit musical literature. Thus, there had been no attempt to systematize the music, and there was a considerable gap between performance and theory until the present century. Vishnu Narayana Bhatkande, one of the leading Indian musicologists of this century, contributed a great deal toward diminishing the gap. Being both a scholar and a performer, he devoted much effort to collecting and notating representative versions of a number of ragas from musicians belonging to different family traditions, or *gharānā*s. Based on this collection, he concluded that most of the ragas of North Indian music can be grouped into the following scales, called *thāṭa*s (compare the South Indian *mela*s shown above):

kalyāṇa	C	D	E	F♯	G	A	B	c
bilāvala	C	D	E	F	G	A	B	c
khamāja	C	D	E	F	G	A	B♭	c
bhairava	C	D♭	E	F	G	A♭	B	c
pūrvī	C	D♭	E	F♯	G	A♭	B	c
mārvā	C	D♭	E	F♯	G	A	B	c
kāfī	C	D	E♭	F	G	A	B♭	c
āsāvarī	C	D	E♭	F	G	A♭	B♭	c
bhairavī	C	D♭	E♭	F	G	A♭	B♭	c
toṛī	C	D♭	E♭	F♯	G	A♭	B	c

The *thāṭa*s do not cover all the ragas used in North Indian music, but there is reason to believe that most of the ragas having scales other than the above are relatively modern innovations. New ragas are constantly being created, and some North Indian musicians are using the vast potential of the South Indian *mela* system as their source of inspiration.

Scale classifications as opposed to raga

Mela and *thāṭa* are theoretical devices for the classification of ragas. Ragas have scalar elements, such as specified ascending and descending movements, that might or might not employ adjacent steps. They may also employ oblique or zigzag movements. Ragas can be heptatonic, hexatonic, or pentatonic and may also have accidentals (sharpened or flattened notes) that occur only in specific melodic contexts. A further distinction between scale and raga is found in the varying emphasis placed on different notes in a raga. Ragas, furthermore, also have melodic elements, such as certain recurrent nuclear motives (brief melodic fragments) that enable the raga to be identified more easily. One scale type can be the basis for perhaps 20 or 30 ragas, in which case it is the nonscalar elements that provide the distinguishing features of each raga in the group.

Rhythmic organization. *South India.* Just as the system of classifying raga is better organized in South Indian music, so too is the system of classifying *tāla,* or time measure. The main group is composed of 35 *tāla*s, called the *sūlādi-tāla*s. Each *tāla* is composed of one, two, or three different units: short, medium, and long. The medium unit is twice the duration of the short; the long unit is, however, a variable and may be three, four, five, seven, or nine times the duration of the short. There are seven basic *tāla* patterns, and, because the long unit of these *tāla*s can be of five different durations, the total number of *tāla*s in this system is 35. The basic *tāla* patterns are:

dhruva-tāla—long, medium, long, long
maṭhya-tāla—long, medium, long
rūpaka-tāla—medium, long
jhampā-tāla—long, short, medium
tripuṭa-tāla—long, medium, medium
āṭa-tāla—long, long, medium, medium
eka-tāla—only a single long.

The total duration of each pattern is controlled by the duration of the variable long; thus, if the long unit is five times the short, a *tāla* pattern such as *dhruva-tāla* will be 5 + 2 + 5 + 5, or 17 units. Several of these *tāla*s have the same total duration but are distinguished from each other by their internal subdivisions. In the course of a performance, the vocalist, as well as the audience, may mark the time by clapping, hand waving, and finger counting.

In addition to the *sūlādi-tāla*s, there are four *cāpu-tāla*s that are used in South Indian classical music. Said to derive from folk music, they consist of two sections of unequal length, 1 + 2, 2 + 3, 3 + 4, and 4 + 5. Of these, the 3 + 4 combination is the most prominent. On rare occasions a performer may use one of the "classical" *tāla*s referred to in Sanskrit texts. These generally involve long time cycles composed of as many as 100 short units. The most frequently heard time measures, however, are *ādi-tāla,* a modified eight-beat version of *tripuṭa-tāla* (4 + 2 + 2); *miśra-cāpu-tāla* (3 + 4); and *rūpaka-tāla* (4 + 2). The difficult and long *tāla*s are used primarily as a tour de force. Each *tāla* may be performed in either slow, medium, or quick tempo; there is no gradual acceleration as in North Indian music.

North India. In North Indian music the *tāla*s are fewer and not organized in any systematic manner. As in South Indian music, the two main factors are the duration of the time cycle and the subdivisions within the cycle. Each of these subdivisions is marked by a clap or a wave, with the greatest emphasis falling on beat 1 of the cycle, which is called *sam.* North Indian *tāla*s have a further feature, the *khālī* ("empty"), a conscious negation of stress occurring at one or more points in each *tāla* where one would expect a beat. It often falls at the halfway point in the time cycle and is marked by a wave of the hand. There is nothing comparable to the *khālī* in the South Indian system. A further distinguishing feature found only in North Indian *tāla*s is the emphasis placed on the characteristic drum pattern of each *tāla,* called *ṭhekā.* Two *tāla*s might have the same duration and subdivisions but might, nevertheless, be differentiated from each other by different characteristic drum patterns. In addition, the *tāla*s are also associated with different forms of song and even particular tempi. The usual North Indian *tāla*s range from six to 16 time units in duration. The most popular are *tīn-tāla* (4 + 4 + 4 + 4), *eka-tāla* (2 + 2 + 2 + 2 + 2 + 2), *jhap-tāla* (2 + 3 + 2 + 3), *kaharavā* (4 + 4), *rūpaka-tāla* (3 + 2 + 2), and *dādrā* (3 + 3). *Tīn-tāla* should not be confused with Western $\frac{4}{4}$, or common time, for the time cycle repeats only after 16 units and is more like four bars of common time.

Characteristic North Indian rhythmic features

Musical forms and instruments. *South India.* Both raga and *tāla* provide bases for composition and improvisation in Indian classical music. A performance usually begins with an improvised section, called *ālāpa,* played in free time without accompaniment of drums. It may have various sections and might on occasion last half an hour or longer. It is followed by a composed piece in the same raga, set in a particular *tāla.* In South Indian music all composed pieces are primarily for the voice and have lyrics. In North India, however, there are also some purely instrumental compositions, called *gat* and *dhun.* The emphasis on the composition varies in the different forms of song and, to some extent, in the interpretation of the performer. In South Indian music the composed piece is generally emphasized more than in the North. Much of the South Indian repertoire of compositions stems from three composers, Tyagaraja, Muthuswami Dikshitar, and Syama Sastri, contemporaries who lived in the second half of the 18th and the beginning of the 19th centuries. The devotional songs that they composed, called *kṛti,* are a delicate blend of text, melody, and rhythm and are the most popular items of a South Indian concert. The composed elements in these songs sometimes include sections such as *niraval,* melodic variations with the same text, and *svara-kalpana,* passages using the Indian equivalent of the sol-fa syllables, which are otherwise improvised.

The longest item in the South Indian concert, called *rāgam-tānam-pallavi,* is, on the other hand, mostly improvised. It begins with a long *ālāpa,* called *rāgam* in

Interaction of composition and improvisation

this context, presumably because this elaborate, gradually developing *ālāpa* is intended to display the raga being performed in as complete a manner as possible, without the limitations imposed by a fixed time measure. This is followed by another improvised section, *tānam,* in which the singer uses meaningless words to produce more or less regular rhythms, but still without reference to time measure. This section, too, is without drum accompaniment. The final section, *pallavi,* is a composition of words and melody set in a particular *tāla,* usually a long or complex one. The *pallavi* may have been composed by the performer himself and be unfamiliar to his accompanists, usually a violinist who echoes the singer's phrases and a drummer who plays the mridanga, a double-ended drum. The statement of the composition is followed by elaborate rhythmic and melodic variations that the accompanists are expected to follow. It is customary to have a drum solo at the end of the *pallavi,* and the performance concludes with a brief restatement of the *pallavi.*

Other forms used in South Indian classical music derive largely from the musical repertoire of *bhārata-nāṭyam,* the classical South Indian dance. The *varṇam,* a completely composed piece, serves mainly as a warming up and is performed at the beginning of a concert. *Pada* and *jāvali* are two kinds of love songs using the poetic imagery characteristic of the romantic-devotional movement mentioned earlier. *Tillānā* has a text composed mostly of meaningless syllables, which may include the onomatopoeic syllables used to represent the different drum sounds. This is a very rhythmic piece and is usually sung in fast tempo.

Melodic, drone, and rhythmic instruments

The ensemble used in present-day South Indian classical music consists of a singer or a main melody instrument, a secondary melody instrument, one or more rhythmic percussion instruments, and one or more drone instruments. The most commonly heard main melody instruments are the *vīṇā,* a long-necked, fretted, plucked lute with seven strings; the *veṇu,* a side-blown bamboo flute; the *nagaswaram,* a long, oboe-like, double-reed instrument with finger holes; the violin, imported from the West about 200 years ago, played while seated on the floor with the scroll resting on the player's left foot; and the *goṭṭuvādyam,* a long-necked lute without frets, played like the Hawaiian guitar, with a sliding stop in the left hand.

The violin is by far the most common secondary melody instrument in South India. It plays in unison where the passage is composed but imitates the voice or main melody instrument in the improvised passages. Of the rhythm instruments, the mridanga, a double-conical, two-headed drum, is the most common. Others include the *kañjīrā,* a tambourine; the *ghaṭam,* an earthenware pot without skin covering; the *morsing,* a metallic jew's harp; and the *tavil,* a slightly barrel-shaped, double-ended drum, which accompanies the *nagaswaram.* The most prominent drone instrument is the four-stringed tamboura, a long-necked lute without frets. It accompanies the voice and all melody instruments, except the *nagaswaram,* which is usually accompanied by the *oṭṭu,* a longer version of the *nagaswaram* but without finger holes. A hand-pumped harmonium drone, called *śruti* or *śruti* box, sometimes replaces the *oṭṭu* or the tamboura.

North India. The most common vocal form in North Indian classical music at the present time is the *khyāl,* a Muslim word meaning "imagination." The *khyāl* is contrasted with the *dhruvapada* (now known as *dhrupad*), which means "fixed words." The two forms existed side by side in the Islāmic period, and it is only in the last century or two that *khyāl* has achieved ascendancy. There are two types of *khyāl.* The first is sung in extremely slow tempo, with each syllable of the text having extensive melisma (prolongation of a syllable over many notes), so that the words are virtually unrecognizable. It is not usually preceded by a lengthy *ālāpa;* instead, *ālāpa*-like phrases are generally sung against the very slow time measure to the accompaniment of the drums. Also characteristic of the *khyāl* are the *sargam tānas,* passages using the Indian equivalent of the sol–fa syllables, and the *ā-kār tānas,* which are rapid runs sung to the syllable *aah.* The second type of *khyāl,* which may be as much as eight times faster than the slow and is generally set in a different *tāla,* follows the slow. Its composed portion is usually quite short, and the main features of the improvisation are the *ā-kār tānas.* Occasionally, a composition called *tarānā,* made up of meaningless syllables, may replace the fast-tempo *khyāl.*

The *ṭhumrī* is another North Indian vocal form and is based on the romantic-devotional literature inspired by the *bhakti* movement. The text is usually derived from the Rādha-Krishna theme and is of primary importance. The words are strictly adhered to, and the singer attempts to interpret them with his melodic improvisations. It is quite usual for a singer to deviate momentarily from the raga in which the composition is set, by using accidentals and evoking other ragas that might be suggested by the words, but he always returns to the original raga.

Some of the North Indian musical forms are very like the South Indian. The vocal forms *dhrupad* and *dhamār* resemble the *rāgam-tānam-pallavi.* They begin with an elaborate *ālāpa* followed by the more rhythmic but unmeasured *non-tom* using meaningless syllables such as *te, re, na, nom,* and *tom.* Then follow the four composed sections of the *dhrupad* or *dhamār,* the latter being named after *dhamār-tāla* of 14 units (5 + 5 + 4) in which it is composed, the former name derived from *dhruvapada.* The song, usually in slow or medium tempo, is first sung as composed; then the performer introduces variations, the words often being distorted and serving merely as a vehicle for the melodic and rhythmic improvisations. Although the *dhrupad-dhamār* form has been out of favour for over a century, it is now apparently being revived.

Instrumental forms

Instrumental music has gained considerable prominence in North India in recent times. The most common instrumental form is the *gat,* which seems to have derived its elements from both *dhrupad* and *khyāl.* It is usually preceded by *ālāpa* and *jor,* which resemble the *ālāpa* and *non-tom* sections of the *dhrupad.* On plucked stringed instruments these two movements are often followed by *jhālā,* a fast section in which the rhythmic plucking of the drone strings is used to achieve a climax. The performer usually pauses before the composed *gat* is introduced. Like the *khyāl,* the *gat* can be in slow or fast tempo. The composition is generally short, and the emphasis is on the improvisations of the melody instrumentalist and the drummer, who for the most part alternate in their extemporizing. The final climax may once again be achieved by a *jhālā* section, in which the tempo is accelerated quite considerably. Other forms played on instruments are the *ṭhumrī,* basically an instrumental rendering of a vocal *ṭhumrī,* and *dhun,* which is derived from a folk tune and does not usually follow a conventional raga. One may also hear a piece called *rāga-māla* (literally, "a garland of ragas"), in which the musician modulates from one raga to another, finally concluding with a return to the original raga.

Musical instruments of North India

The most prominent melody instruments used in North Indian classical music are the sitar, a long-necked fretted lute; *surbahār,* a larger version of the sitar; the sarod, a plucked lute without frets and a shorter neck than that of the sitar; the *sāraṅgī,* a short-necked bowed lute; the bansuri, a side-blown bamboo flute with six or seven finger holes; the *sheh'nai,* a double-reed wind instrument similar to the oboe, but without keys; and the violin, played in the same manner as in South India. Secondary melody instruments are used only in vocal music, the two most common being the *sāraṅgī* and the keyboard harmonium, an import from the West. The violin and the *surmaṇḍal,* a plucked board zither, are also used in this context. In recent times, instrumental duets, in which the musicians improvise alternately, have grown in popularity. In these duets the musicians may imitate each other's phrases, temporarily creating something of the effect of a secondary melody instrument.

As with South Indian music, the drone is usually provided by a tamboura (Bengali *tanpura*) or a hand-pumped reed drone similar to the harmonium but without a keyboard, called *sur-peṭī* in North India. The *sheh'nai* is usually accompanied by one or more drone *shehnais,* called *sur.*

The rhythmic accompaniment is usually provided on the tabla, a pair of small drums played with the fingers. As

accompaniment to the somewhat archaic *dhrupad*, however, the pakhavāj, a double-conical drum, similar to the South Indian mridanga, is generally used. The *sheh'nai* in classical music is usually accompanied by a small pair of kettledrums, called *ḍukaṛ-ṭikaṛ*.

Interaction with Western music. It is in the sphere of musical instruments that the influence of Western music is most obvious. In addition to the violin and the harmonium, many other Western instruments are occasionally used. These include the clarinet, saxophone, trumpet, guitar, mandolin, and organ. Scholars have criticized the use of some of these instruments on the ground that their tuning, being based on the Western tempered scale (having 12 equal semitones), is not suitable for the performance of Indian music, and All-India Radio has forbidden the use of the harmonium in its programs. Most of the leading North Indian singers, however, have been using the harmonium as a secondary melody instrument for many years and have continued to do so in concerts and on recordings.

Apart from the area of musical instruments, Indian music appears to have absorbed very little of Western music. It is possible that some modern developments in classical music might have been inspired by Western music. These include the slightly increased use of chromaticism (using a succession of semitones) and some of the new drone tunings in which the major third is added (making for example, the drone on the first, third, and fifth notes of the scale, rather than on the first and fifth only). But the evidence is not conclusive, and it could equally be argued that these are natural developments within the system. Western technology has, of course, had a profound influence on Indian music. Sound-amplification devices have made concerts available to large audiences, and the intimate atmosphere in which the music was traditionally performed is now seldom encountered. The Indian musician has been obliged to adapt his music, once played before a select and musically educated group of listeners, to new circumstances involving a mass of people, many of whom are unable to appreciate the finer points of the music. The use of microphones during concerts has had a marked effect on voice production, and, since the voice no longer needs to project over distances, many modern singers now sing with a relaxed throat and produce a more mellow tone.

Impact outside of India

Since the mid-1950s, Indian classical music has been performed fairly regularly in the West. Initially, the audiences were composed mainly of South Asians, but gradually an increasing number of Westerners have been attending the concerts. Perhaps the music would not have reached beyond a very limited audience were it not for the interest shown by the American violinist Yehudi Menuhin, who sponsored a number of programs in the West, and the British popular-music group the Beatles, who attempted to incorporate the sound of the sitar and other elements of Indian culture into the world of Western popular music. At the same time, several North Indian instrumentalists, such as Ravi Shankar, Ali Akbar Khan, Vilayat Khan, Imrat Khan, and Nikhil Banerjee, were received with overwhelming enthusiasm by Western audiences. By the end of the 1960s the sitar and tabla were heard frequently in Western pop music, jazz, cinema, and television programs, as well as in radio and television advertisements.

Within three or four years, the mass Western involvement with Indian music was over. It is perhaps too early to assess the full impact of this period. It would seem that Indian music has not as yet had any significant influence on Western music. A few modern composers have attempted to incorporate elements of Indian music into their compositions, but their works remain as experiments. The fusing of Indian music and jazz would seem to have more possibilities, since improvisation is an important factor in both, but present attempts have not fulfilled this expectation. Most of these attempts seem to be premature and based on an inadequate understanding of one or the other musical system. (N.A.J.)

Dance and theatre

Theatre and dance in South Asia stem principally from Indian tradition. The principles of aesthetics and gesture language in the *Nāṭya-śāstra*, a 2,000-year-old Sanskrit treatise on dramaturgy, have been the mainstay of all the traditional dancers and actors in India. Even folk performers follow some of its conventions; *e.g.*, the *Kandyan* dancers of Ceylon (now Sri Lanka), who preserve some of the whirls and spins described in this ancient Indian text. Despite the influence of the different religious waves that swept the subcontinent through the centuries, the forms of South Asian dance and theatre were always able to preserve their ancient core.

Traditionally, dance and acting are inseparable. The classical South Asian dancer, equipped with a repertoire of gesture language, alternates between *nṛtta*, pure dance; *nṛtya*, interpretive dance; and *nāṭya*, dance with a dramatic element. (The Sanskrit word *naṭa* means a dancer-actor.) Traditional theatre throughout both South and Southeast Asia is a combination of music, dance, mime, stylized speech, and spectacle. The classical and folk actor must be a dancer, a singer, and a mime in one.

Influence on Southeast Asia

Between the 2nd century BC and the 8th century AD, South Indian kings sent overseas trade missions, priests, court dancers, and sometimes armies to Southeast Asia. During these years of cultural expansion, Indian dance forms, mythological lore, and the language of gesture flourished in Burma, Cambodia, Java, Sumatra, and Bali. Later, when India's economic and political power shrank, its cultural empire remained intact. Even when these Southeast Asian countries embraced Buddhism or Islām, they continued performing dance dramas with Hindu gods and goddesses, adding to these their own local myths, costumes, and masks. The two Hindu epics, the *Rāmāyaṇa* and the *Mahābhārata*, storehouses of dramatic personae of traditional dramas, have been absorbed by these countries as part of their own cultural heritage. Some dance forms and gesture vocabulary that died out in their land of birth have been preserved in Bali. For a discussion of the dance and theatre of Southeast Asia see the article SOUTHEAST ASIAN ARTS.

THE PERFORMING ARTS IN INDIA

The royal courts and temples of India traditionally have been the chief centres of the performing arts. In ancient times, Sanskrit dramas were staged at seasonal festivals or to celebrate special events. Some kings were themselves playwrights; the most notable of the playwright-kings was Śūdraka, the supposed 4th-century author of *Mṛcchakaṭika* (The Little Clay Cart). Other well-known royal dramatists include Harṣa, who wrote *Ratnāvālī* in the 7th century; Mahendravikramavarman, author of the 7th-century play *Bhagavad-Ajjukiya;* and Viśākhadatta, creator of the 9th-century drama *Mudrārākṣasa*.

In the 4th century BC, Kauṭilya, the chief minister of Emperor Candragupta, referred in his book on the art of government, the *Artha-śāstra*, to the low morals of players and advised the municipal authorities not to build houses in the midst of their villages for actors, acrobats, and mummers. But, in the glorious era of the Hindu kings during the first eight centuries after Christ, actors and dancers were given special places of distinction. This tradition continued in the princely courts of India even under British rule. *Kathākali* dance-drama, for instance, was created by the Raja of Kottarakkara, ruler of one of the states of South India in the 17th century. The powerful peshwas of the Marāthā kingdom in the 18th century patronized the *tamasha* folk theatre. Nawab Wajid Ali Shah (flourished mid-19th-century) was an expert *kathak* dancer and producer of Krishnalore plays in which his palace maids danced as the *gopīs* (milkmaids who were devotees of Krishna). Maharajas of Travancore and Mysore competed with each other for the excellence of their dance troupes. In the 20th century, the Maharaja of Banaras (Vārānasi) carried on this tradition by being patron and producer of the spectacular *rāmlīlā*, a 31-day cycle play on Rāma's life that he witnessed every night while sitting on his royal elephant. On special nights the spectators numbered more than 30,000.

Royal patronage

Dance is a part of all Hindu rituals. Farmers dance for a plentiful harvest, hunters for a rich bag, fishermen for a good catch. Seasonal festivals, religious fairs, mar-

riages, and births are celebrated by community dancing. A warrior dances before the image of his goddess and receives her blessings before he leaves for battle. A temple girl dances to please her god. The gods dance in joy, in anger, in triumph. The world itself was created by the Cosmic Dance of Lord Śiva, who is called Naṭarāja, the king of dancers, and worshipped by actors and dancers as their patron.

Religious festivals

Religious festivals are still the most important occasions for dance and theatrical activity. The *rāmlīlā kṛṣṇalīlā* and *rāslīlā* in North India (Uttar Pradesh, Delhi, Rājasthān, Haryana, and Punjab), the *chhau* masked dance-drama in Saraikela region in Bihār, and the *bhagavatha mela* in Melatur village in Tamil Nadu are performed annually to celebrate the glory of their particular deities. During the Daśaharā festival every village in North India enacts for a fortnight the story of Rāma's life, with songs, dances and pageants. The *jātrā* in West Bengal is a year-round dramatic activity, but the number of troupes swells to many thousands in Calcutta during the Pūjā festival. The hill and tribal people dance all night to celebrate their community festivals and weddings rich in masks, pageants, and carnivals. In far-flung areas of South Asia, people may not have seen a drama, but there will be hardly a person who has not witnessed or taken part in a community dance.

Audience participation

In folk theatre, traditional dance, classical music, and poetical symposia (especially the Urdu *mushā'irah*), performances are held in the open air or in a well-lit canopied courtyard so that the players can see the spectators and be motivated by their reactions.

For the usually all-night folk dramas, people come with their children, straw mats, and snacks, making themselves at home. At these performances there is a constant inflow and outflow of spectators. Some go to sleep, asking their neighbours to awaken them for favourite scenes. Stalls selling betel leaves, peanuts, and spicy fried things, adorned with flowers and incense and lighted by oil lamps, surround the open-air arena. The clown, an essential character in every folk play, comments on the audience and contemporary events. Zealous spectators offer donations and gifts in appreciation of their favourite actor or dancer, who receives them in the middle of the performance and thanks the donor by singing or dancing a particular piece of his choice. The audience thus constantly throws sparks to the performer, who throws them back. People laugh, weep, sigh, or suddenly fall silent during a moving scene.

In both folk and classical forms of drama, the performer may lengthen or shorten his piece according to audience response. During a *kathak* dance, the drummer, in order to test the perfection of the dancer, disguises the main beat of his drum by slurs and offbeats, a secret he shares with the audience and announces by a loud thump that is synchronized with the dancer's stamping of the foot. At this point in the dance the spectators shout, swaying their heads in admiration. They show their approval and disapproval through delighted groans or sullen headshakes as the performance goes on. In the *rāslīlā*, the audience joins in singing the refrain and marks the beat by hand clapping. At a climactic point the people rock and sway, rhythmically clapping and singing. These practices bind the performers, chanters, and spectators together in a sense of aesthetic pleasure.

Instrumental music and singing

Instrumental music and singing are integral parts of Indian dance and theatre. Musicians, chanters, and drummers sit on the stage in view, a tradition observed throughout almost all of Asia. They watch the dancer and play on their instruments following his movements, whereas in the West the movements of a ballerina are timed and controlled by the already-written music. An Indian dancer is constantly reacting to the accompanying musician, and vice versa. He may signal the chanters and drummers and even instruct them during the performance without spoiling its aesthetic effect.

In some classical dance forms, such as *kuchipudi*, the dancer sings in voiceless whispers as she dances. In *bhārata-nāṭyam* the dance movements are like sculpted music in space, and the accompanying musician is invariably a dance guru (teacher). In *kathak* the rhythmic syllables beaten out by the dancer with her feet are vocalized by the singer and then chirped out by the drummer. No folk dancing is complete without the use of drum and vocal singing. Women's folk singing such as the *giddha* in the Punjab and the men's *kīrtan* in West Bengal takes the form of dance when the rhythm becomes fast.

In folk theatre this relationship is even more apparent. *Rāslīlā* dance sequences are interspersed with the singing as a decorative frill, to accentuate emotional appeal, or to mark the climax of a song. The *yakṣagāna* hero gives a brisk dance number to announce his entry. In many folk forms of opera (*bhavai, terukkūttu,* and *nautanki*), the characters sing and dance at the same time or alternate. Ballad singers from the states of Orissa and Andhra Pradesh dramatize their singing by strong facial gestures and rhythmic ankle bells and execute dance phrases between the narrative singing. On the other hand, no one can imagine a dancer who is not at the same time a musician. This double aesthetic discipline enriches both of these arts, and the Indian audience is conditioned to this tradition.

INDIAN DANCE

Dance in India can be organized into three categories: classical, folk, and modern. Classical dance forms are among the best preserved and oldest practiced in the 20th century. The royal courts, the temples, and the guru to pupil teaching tradition have kept this art alive and unchanged. Folk dancing has remained in rural areas as an expression of the daily work and rituals of village communities. Modern Indian dance, a product of the 20th century, is a creative mixture of the first two forms, with freely improvised movements and rhythms to express the new themes and impulses of contemporary India.

The popularity of dance in 20th-century India can be judged from the fact that there is hardly any Indian motion picture that does not have half a dozen dances in it. In the typical "boy meets girl" film the heroine dances everywhere and anywhere. A film company may not have a script writer (in some cases the financier writes the story himself), but it must have a dance director. To provide ample dance opportunities, motion pictures have been made on the lives of poets, courtesans, and temple dancers and on mythological themes. For these the services of expert dancers are sought.

In the 20th century, classical dance has left the temples and royal courts and is now presented regularly on the stage in large cities. Rich industrialists, international hotels, and the wealthy families of the upper class are the chief patrons. It is not uncommon to have a classical dance recital by a major performer at a business dinner or for the annual function of a club. Some universities have dance as a regular subject in their curricula. Women learn it as a social grace, and young girls learn a few classical dances for greater eligibility in marriage. Folk dancing has also become more common as a contemporary cultural event in the cities. Most colleges have their folk-dance troupes, and even the police of the Punjab have their folk-dance groups to perform the *bhangra*. Folk dance, cut off from its rural settings, has lost much of its original vigour and beauty, but that is the inevitable result of cross-fertilization of regional cultures through folk-troupe exchanges at an interstate level.

Classical dance. *The dance-drama.* India has evolved through its classical and folk traditions a type of dance drama that is a form of total theatre. The actor dances out the story through a complex gesture language, a form that, in its universal appeal, cuts across the multilanguage barrier of the subcontinent. Some of the classical dance-drama forms (*e.g., kathākali, kuchipudi, bhagavatha mela*) enact well-known stories derived from Hindu mythology. The 20th-century dancers Uday Shankar and Shanti Bardhan have created ballets that were inspired by such traditional dance-dramas. Contemporary Indian directors and writers are re-examining traditional dance forms and are using these in their current works for greater psychological appeal and deeper artistic impact. Millions in villages are still entertained by dance-dramas. In spite of the popu-

larity of straight prose plays in the cities, the appeal of dance-drama is unquestionably deeper and more satisfying to the rural Indian, whose aesthetics is still rooted in tradition.

The *Nāṭya-śāstra*

The chief source of classical dance is Bharata Muni's *Nāṭya-śāstra* (1st century BC to 1st century AD), a comprehensive treatise on the origin and function of *nāṭya* (dramatic art that is also dance), on types of plays, gesture language, acting, miming, theatre architecture, production, makeup, costumes, masks, and various *bhāvas* ("emotions") and *rasas* ("sentiments"). No other book of ancient times contains such an exhaustive study of dramaturgy.

Techniques and types of classical dance. According to the *Nāṭya-śāstra,* the dancer-actor communicates the meaning of a play through four kinds of *abhinaya* (histrionic representations): *angika,* transmitting emotion through the stylized movements of parts of the body; *vācika,* speech, song, pitch of vowels, and intonation; *āhārya,* costumes and makeup; and *sāttvika,* the entire psychological resources of the dancer-actor.

The actor is equipped with a complicated repertoire of stylized gestures. Conventionalized movements are prescribed for every part of the body, the eyes and hands being the most important. There are 13 movements of the head, seven of the eyebrows, six for the nose, six for the cheek, seven for the chin, nine for the neck, five for the breasts, and 36 for the eyes. There are 32 movements of feet, 16 on the ground and 16 in the air. Various positions of the feet (strutting, mincing, tromping, splaying, beating, etc.) are carefully worked out. There are 24 single-hand gestures (*asaṃyuta-hasta*) and 13 for combined hands (*saṃyuta-hasta*). One gesture (*hasta*) may mean more than 30 different things quite unrelated to each other. The *patāka* gesture of the hand, for example, in which all the fingers are extended and held close together with the thumb bent, can represent heat, rain, a crowd of men, the night, a forest, a horse, or a flight of birds. The *patāka* hand with the third finger bent (*tripatāka*) can mean a crown, a tree, marriage, fire, a door, or a king. In *karkaṭa* ("crab"), one of the combined hand gestures, the fingers of the hands are interlocked, and this may indicate a honeycomb, yawning after sleep, or a conch shell. Of course, for each of these different meanings, a *hasta* is given a different body posture or action.

The male or female classical dancer portraying a story in a solo performance simultaneously plays two or three principal characters by alternating facial expressions, gestures, and moods. Krishna, his jealous wife Satyabhāmā, and his gentle wife Rukmiṇī, for example, may be played by one person.

Bhāva and *rasa*

The aesthetic pleasure of Hindu dance and theatre is determined by how successful the artist is in expressing a particular emotion (*bhāva*) and evoking the *rasa.* Literally *rasa* means "taste" or "flavour" and is that exalted sentiment or mood that the spectator experiences after witnessing a performance. The critics do not generally concern themselves so much about plot construction or technical perfection of a poem or play as about the *rasa* of a particular work. There are nine *rasa:* erotic, comic, pathetic, furious, heroic, terrible, odious, marvelous, and spiritually peaceful. There are nine corresponding *bhāvas:* love, laughter, pathos, anger, energy, fear, disgust, wonder, and quietude.

Four distinct schools of classical Indian dance—*bhārata-nāṭya, kathākali, kathak,* and *manipuri*—exist in the 20th century, along with two types of temperament—*tāṇḍava,* representing the fearful male energy of Śiva, and *lāsya,* representing the lyrical grace of Śiva's wife Pārvatī. *Bhārata-nāṭya,* which takes its name from Bharata's *Nāṭya-śāstra,* has the *lāsya* character, and its home is Tamil Nadu, in South India. *Kathākali,* a pantomimic dance-drama in the *tāṇḍava* mood with towering headgear and elaborate facial makeup, originated in Kerala. *Kathak* is a mixture of *lāsya* and *tāṇḍava* characterized by intricate footwork and mathematical precision of rhythmic patterns; it flourishes in the north. *Manipuri,* with its swaying and gliding movements, is *lāsya,* and it has been preserved in Manipur state in the Assam Hills. In 1958 the Sangeet Natak Akademi (National Academy of Music, Dance and Drama) in New Delhi bestowed classical status on two other schools of dance—*kuchipudi,* from Andhra Pradesh, and *oṛissī,* from Orissa. These two styles overlap the *bhārata-nāṭya* school and therefore are not as distinctly different in temperament and style as other forms.

The bhārata-nāṭya school. Bhārata-nāṭya (also called *dasi attam*) has survived to the present through the *devadāsīs,* temple dancing girls who devoted their lives to their gods through this medium. Muslim invasions from the north destroyed the powerful Hindu kingdoms in the south but could not disrupt their arts, which took shelter in the temples. After the 16th century, the Muslims overpowered the south completely until the British came, thus giving a setback to Hindu dance. Slowly the institution of *devadāsī* fell into disrepute, and temple dancing girls became synonymous with prostitutes. In the latter half of the 19th century in Tanjore, Chinniyah, Punniah, Vadivelu, and Shivanandam, four talented dancers who were brothers, revived the original purity of *dasi attam* by studying and following the ancient texts and temple friezes, with missing links supplied by the socially spurned *devadāsīs.* Their popularized form of *dasi attam* was called *bhārata-nāṭya.*

A performance of *bhārata-nāṭya* lasts for about two hours and consists of six parts, beginning with *allarippu* (Telegu language, "to decorate with flowers"), a devotional prologue that shows off the elegance and grace of the dancer. The second part is *jātisvaram,* a brilliant blaze of *jātis* ("dance phrases") with *svaras* ("musical sounds"). This is followed by *shabdam,* the singing words that prepare the dancer to interpret through *abhinaya* (gesture language) interspersed with pure dance. The fourth part is *varṇam,* a combination of expressive and pure dance. Then follow the *padams,* songs in Telegu, Tamil, or Kanarese that the dancer dramatizes by facial expressions and hand gestures. The accompanying singer chants the line again and again, and the dancer enacts the clashing and contrasting meanings. Her virtuosity consists of exhausting all possible shades of suggestion. The performance ends with *tillānā,* a pure dance accompanied by meaningless musical syllables chanted to punctuate the rhythm. The dancer explodes into leaps and jumps forward and backward, from right and left, in a state of ecstasy. *Tillānā* ends with three clangs of the cymbals while the dancer executes a triple blaze of *jātis,* thumping her feet with a jingling flourish of ankle bells.

Bhārata-nāṭya has attained world recognition as one of the most exquisite forms of classical dance. Its aspirants go to Tamil Nadu to learn from gurus who still live in villages. Because of its *lāsya* character, performing artists have always been women. But their teachers have invariably been old men who chant the lines to tiny cymbals, controlling the complex rhythm without dancing themselves.

The major 20th-century performers associated with the *bhārata-nāṭya* school of dance are T. Balasaraswathi, especially known for her *abhinaya* (expressive interpretation) of *padams*; Rukmini Devi, who popularized *bhārata-nāṭya* among the upper classes in the 1930s; Yamini Krishnamurthi; and Shanta Rao. Two of the most important gurus were Minakshisundaram Pillai, who injected vigour into *bhārata-nāṭya* by his choreography, and his son-in-law, Chokkalingam Pillai.

The kathākali school. Kathākali (*kathā,* "story"; *kali,* "performance") originated in the 17th century in Kerala, the lush tropical coastal strip of South India washed by the Arabian Sea. It was devised by the Raja of Kottarakkara, who, angry over the refusal of a neighbouring prince to allow his dancers to perform a Sanskrit dance-drama in his court, decided to create his own dance troupe using Malayalam, the spoken language of the people. This school has its own *hastas,* based on a regional text influenced by the *Nāṭya-śāstra* and later treatises. It also has marked elements of energetic ritualistic dances. The makeup has its roots in the grotesque pre-Hindu Dravidian demon masks. Themes are taken mainly from the *Rāmāyaṇa,* the *Śiva-Purāṇa,* the *Bhāgavata-Purāṇa,* the *Mahābhārata,* and other religious texts. The superhuman characters represent primal forces of good and evil at war. Because of its terrifying vigour, men play all the roles.

Indian classical dance.
(Top left) *Bhārata-nātya* traditional dance drama. (Top right) Male and female *kathākali* dancers. (Bottom left) *Kathak* school dancer in Mughal costume. (Bottom right) *Manipuri* style performance of *rās*.
(All but top right) Mohan Khokar. (top right) Foto Features

Characters and costumes

Most *kathākali* characters (except those of women, Brahmins, and sages) wear towering headgear and billowing skirts and have their fingers fitted with long silver nails to accentuate hand gestures. The principal characters are classified into seven types. (1) *Pacca* ("green") is the noble hero whose face is painted bright green and framed in a white bow-shaped sweep from ears to chin. Heroes such as Rāma, Lakṣmaṇa, Krishna, Arjuna, and Yudhiṣṭhira fall into this category. (2) *Katti* ("knife"), haughty and arrogant but learned and of exalted character, has a fiery upcurled moustache with silver piping and a white mushroom knob at the tip of his nose. Two walrus tusks protrude from the corners of his mouth, his headgear is opulent, and his skirt is full. Duryodhana, Rāvaṇa, and Kichaka belong to this type. (3) *Chokannatadi* ("red beard"), power-drunk and vicious, is painted jet black from the nostrils upward. On both cheeks semicircular strips of white paper run from the upper lip to the eyes. He has black lips, white warts on nose and forehead, two long curved teeth, spiky silver claws, and a blood-red beard. (4) *Velupputadi* ("white beard") represents Hanuman, son of the wind god. The upper half of his face is black and the lower red, marked by a tracery of curling white lines. The lips are black, the nose is green, black squares frame the eyes, and two red spots decorate the forehead. A feathery gray beard, a large furry coat, and bell-shaped headgear give the illusion of a monkey. (5) *Karupputadi* ("black beard") is a hunter or forest dweller. His face is coal black with crisscross lines drawn around the eyes. A white flower sits on his nose, and peacock feathers closely woven into a cylinder rise above his head. He carries a bow, quiver, and sword. (6) *Kari* ("black") is intended to be disgusting and gruesome. Witches and ogresses, who fall into this category, have black faces marked with queer patterns in white and huge, bulging breasts. (7) *Minnukku* ("softly shaded") represents sages, Brahmins, and women. The men wear white or orange dhotis. Women have their faces painted light yellow and sprinkled with mica, and their heads are covered by saris.

Under a flower-decked canopy on a square, ground-level stage a tall brass worship lamp brimming with coconut oil burns brightly. The musicians and dancers bow before it before they start performing. Drummers standing in one corner pound the *cenda*, a barrel-shaped drum with a piercing, clattering sound suited for battle scenes, and continue throughout the performance, almost without respite.

Two men hold a 12-foot by six-foot (four-metre by two-metre) embroidered hand curtain from behind which the principal characters make their entrances. They dance, grab the trembling curtain, and give vivid facial expressions with fearful glances and grunts. This "peering over the curtain," called *tiranokku,* is a close-up that offers an actor full scope to display his art. At a climactic moment the curtain is whisked away and the character enters in full splendour. The performance lasts all night, the singers singing the text that the dancers act out in an elaborate gesture language.

Well-known performers of *kathākali* include Guru Chandu Panikkar, Guru Kunju Kurup, Ramunni Nair, and Kalamandalam Krishna Nair. The dancers Guru Gopi Nath and Krishnan Kutty have both emphasized simplification of the use of towering headgear and thick-crusted, elaborate makeup, so that the art may be more commonly understood.

The kathak school. *Kathak,* born of the marriage of Hindu and Muslim cultures, flourished in North India under Mughal influence. *Kathak* dancers retain their 17th-century costumes but are steeped in Rādhā and Krishna love lore. Krishna, playing his flute in the Vṛndāvana woods on the bank of the Yamuna River, is surrounded by the *gopīs* ("milkmaids"). Their play is the eternal game of the god and his devotees, the hide-and-seek of man and woman. This spiritual relationship is deeply passionate, with erotic love-play. Slowly the dance degenerated and found shelter in bawdy houses, where nautch girls practiced the art to make themselves more tantalizing. In the beginning of the 20th century it was reclaimed and revived, however, mainly through the efforts of Kalkaprasad Maharaj, whose three sons Achchan, Lachchu, and Shambhu, perfected the art.

Because of its mixed *lāsya* and *tāṇḍava* temperament, *kathak* is popular with both females and males. In *bhāratanāṭya,* footwork is synchronized with hand gestures and eye movements, but *kathak* has no such rigid technique. It takes its movements from life, stylizes them, and adds complex rhythmic patterns. The mathematical precision in doubling and quadrupling the beat with quick transfers and shifts makes the onlookers dizzy.

Kathak performers

A female *kathak* dancer generally wears a brocade blouse, a long, wide, shimmering silk skirt, a transparent tissue scarf of gold threads, and a heavy cluster of ankle bells. A musician, generally the guru, sits beside the drummer on the floor and vocalizes the complicated syllables of the drum that the dancer beats out with her feet. *Kathak's* basic dance posture and some of the steps can be traced to the *rāslīlā* of Braj Bhoomi. The musical refrain, which is called *lehra,* provides the base on which the drummer and the dancer execute a rich tapestry of rhythmic patterns. Beats are called *mātrās* and the footwork *tatkar.* Important elements of the dance are *chakkars, torahs,* and *tihais. Chakkar* denotes whirling with great speed and stopping for a fraction of time after each whirl within the prescribed beat while at the same time maintaining the beauty of the form. *Torah* is a composition consisting of rhythmic syllables. *Tihai* is the repetition of a phrase of rhythmic syllables used to adorn the concluding part of a *torah.* There are two styles of *kathak:* Jaipur *gharana* and Lucknow *gharana.* While the Lucknow *gharana* excels in *bhāva,* the Jaipur *gharana* specializes in brilliance of footwork.

In the 20th century the major performers of *kathak* include Shambhu Maharaj, who specialized in *bhavapradarśan* ("display of emotion"), and Sunder Prasad, who concentrated on the *tala* and *layakari* aspects of the dance. Birju Maharaj, Gopi Krishan, Sitara Devi, and Damayanti Joshi all have important reputations in India as well as abroad.

The manipuri school. *Manipuri* has survived in the sheltered valley of Manipur in the Assam Hills. It remained aloof not only from foreign influences but also from the main Indian trends. Its isolation was broken only in the 1920s, when Rabindranath Tagore visited the valley and invited a leading guru of the area, Atomba Singh, to teach at his school in Santiniketan. The supple movements of *manipuri* dance were suitable for Tagore's lyrical dramas, and he therefore employed them in his plays and introduced the dance as a part of the curriculum at his institution.

The *manipuri* dancer wears a large, stiff skirt that is glittering with round mirror pieces and a shimmering gauze veil. Her hair is done up in a high rolled crown that is adorned with chains of white blossoms, and her luminous cheeks and forehead are decorated with dots of sandalwood paste.

Style of the *manipuri* dance

Known for its femininity, *manipuri* is marked by a slow, swooning rhythm. The dancer, with her hips thrust back and head tilted on one side, turns and sways and glides as if in a dream. The immobility of her face, like that of a mask, is in sharp contrast with the other three schools of dance, in which the face and eyes are a major source of expression.

The *manipuri* drummer, his naked torso in a white dhoti with a red border tucked up above his knees, dances while he plays on the drum. He slaps and thumps; the drum rumbles and howls and chuckles. Drunk with its rhythm, the drummer dances in wild, frenzied leaps. His energetic and electric movements are a masculine counterpart to the slow, undulating patterns woven by the female dancer.

Chief 20th-century exponents of *manipuri* include Atomba Singh, who preserved the tradition of *rās* dancing, and Amubi Singh.

The kuchipudi school. *Kuchipudi* dance-dramas owe their origin to the small village of Kuchipudi (Kuchelapuram) in Andhra Pradesh. Their form was originated in the 17th century by Sidhyendra Yogi, creator of the superb dance-drama *Bhama Kalapam,* which is the story of charming Satyabhāma, jealous wife of Lord Krishna. Sidhyendra Yogi taught the art to Brahmin boys of Kuchipudi and gave a performance with them in 1675 for the Nawab of Golconda, who was so pleased that he granted Kuchipudi to the Brahmin Bhavathas for the preservation of this art. Even in the 20th century every Brahmin of Kuchipudi is expected to perform at least once in his life the role of Satyabhāma as an offering to Lord Krishna.

The *kuchipudi* dance begins with worship rituals. A male dancer moves about sprinkling holy water, and then incense is burned. *Indra-dhvaja* (the flagstaff of the god Indra) is planted on the stage to guard the performance against outside interference. Women sing and dance with worship lamps, followed by the worship of Gaṇeśa, the elephant god, who is traditionally petitioned for success before all enterprises. The *bhagavatha* (stage manager-singer) sings invocations to the goddesses Sarasvatī (Learning), Lakṣmī (Wealth), and Paraśakti (Parent Energy), in between chanting drum syllables.

Two men hold up the traditional coloured curtain. A long gold-embroidered braid is hung on the curtain as a challenge to anyone among the spectators who dares to act and dance. If anyone should take up this braid, the hero playing the female character Satyabhāma will cut off "her" hair. The principal characters are introduced from behind the curtain after each one has done a brisk dance, and at that time the *bhagavatha* sings out the background and function of each. All roles are traditionally played by men (but in recent times by women also), and all the four elements of *abhinaya* are used—dance, song, costume, and psychological resources. Thus, *kuchipudi* differs from other classical dances in which the performers do not sing.

Among the major *kuchipudi* dancers of the 20th century are Guru Chinta Krishnamurthi, Vedantam Satyanarayana, and Yamini Krishnamurthi.

Odissi. *Odissi,* practiced in Orissa, claims to be over 2,000 years old and the true inheritor of the *Nāṭya-śāstra* tradition. It originated and was initially developed in the temples and later flourished in the courts as well. Many of the 108 basic dance units (*karaṇas*) mentioned in the *Nāṭya-śāstra* can be found only in *odissi,* and many of its dance poses are sculpted on the exterior of the temples of Bhubaneswar, Konārak, and Purī. Kelu Charan Mahapatra and Indrani Rehman are the principal 20th-century figures associated with *odissi.*

Other classical dance forms. Among other classical or

Indian folk dance.
(Left) *Chhau* dance of Bihār showing boatman and wife. (Centre) *Kacchi ghori* dancers of Rājasthān. (Top right) *Ghoomar* dancers of Rājasthān. (Bottom) *Garabā* dancers of Gujarāt.
(Left, bottom) Mohan Khokar, (centre, top right) Foto Features

Bhagavatha mela, mohini attam, and *kuṟavañci*

semiclassical dance forms are *bhagavatha mela, mohini attam,* and *kuravañci.* Performed at the annual Narasimha Jayanti festival in Melatur village in Tamil Nadu, the *bhagavatha mela* uses classical gesture language with densely textured Karnatak music. Its repertoire was enriched by the musician-poet Venkatarama Sastri (1759–1847), who composed important dance-dramas in the Telugu language. *Mohini attam* is based on the legend of the Hindu mythological seductress Mohinī, who tempted Śiva. It is patterned on *bhārata-nāṭya* with elements of *kathākali.* It uses Malayalam songs with Karnatak music. *Kuṟavañci* is a dance-drama of lyrical beauty prevalent in Tamil Nadu. It is performed by four to eight women, with a gypsy fortune-teller as initiator of the story of a lady pining for her lover. Formally, it is a mixture of the folk and classical types of Indian dance.

Folk dance. Indian folk dances have an inexhaustible variety of forms and rhythms. They differ according to region, occupation, and caste. The half-naked Adivasis (aboriginal tribes) of central and eastern India (Murias, Bhils, Gonds, Juangs, and Santāls) are the most uninhibited in their dancing. There is hardly a national fair or festival where these dances are not performed. The most impressive occasion occurs every January 26 on Republic Day, when dancers from all parts of India come to New Delhi to dance in the vast arena of the National Stadium and along a five-mile parade route.

Types of folk dances

It is difficult to categorize Indian folk dances, but generally they fall into four groups: social (concerned with such labours as tilling, sowing, fishing, and hunting); religious; ritualistic (to propitiate an angry goddess or demon with magical rites); masked (a type that appears in all the above categories).

The *kolyacha* is among the better known examples of social folk dance. A fisherman's dance indigenous to the Konkan coast of western central India, the *kolyacha* is an enactment of the rowing of a boat. Women wave handkerchiefs to their male partners, who move with sliding steps. For wedding parties young Kolis dance in the streets carrying household utensils for the newlywed couple, who join the dance at its climax.

The national social folk dance of Rājasthān is the *ghoomar,* danced by women in long full skirts and colourful *chuneris* (squares of cloth draping head and shoulders and tucked in front at the waist). Especially spectacular are the *kacchi ghori* dancers of this region. Equipped with shields and long swords, the upper part of their bodies each arrayed in the traditional attire of a bridegroom and the lower part concealed by a brilliant-coloured papier-mâché horse built up on a bamboo frame, they enact jousting contests at marriages and festivals. Bawaris, by tradition a criminal tribe, generally are expert in this form of folk dance.

In the Punjab, the most electrifying social folk dance is the male harvest dance, *bhangra,* which is also popular in the Punjab province of Pakistan. This dance is always punctuated by a song. At the end of every line the drum thunders. The last line is taken up by all the dancers in a chorus. In ecstasy they spring, bellow, shout, and gallop in a circle, madly wiggling their shoulders and hips. Any man of any age can join.

The Lambadi Gypsy women of Andhra Pradesh wear mirror-speckled headdresses and skirts and cover their arms with broad, white bone bracelets. They dance in slow, swaying movements, with men acting as singers and drummers. Their social dance is imbued with impassioned grace and lyricism and is less wild than that of Gypsies in other parts of the world.

The bison-horn dance of the Muria tribe in Madhya Pradesh is performed by both men and women, who traditionally have lived on equal terms. The men wear a horned headdress with a tall tuft of feathers and a fringe of cowry shells dangling over their faces. A drum shaped like a log is slung around their necks. The women, their heads surmounted by broad, solid-brass chaplets and their breasts covered with heavy metal necklaces, carry sticks in their right hands like drum majorettes. Fifty to 100 men and women dance at a time. The male "bisons" attack and fight each other, spearing up leaves with their horns and chasing the female dancers in a dynamic interpretation of nature's mating season.

The Juang tribe in Orissa performs bird and animal

dances with vivid miming and powerful muscular agility.

Religious folk dance

Some major examples of religious folk dances are the *dindi* and *kala* dances of Mahārāshtra, which are expressions of religious ecstasy. The dancers revolve in a circle, beating short sticks (*dindis*) to keep time with the chorus leader and a drummer in the middle. As the rhythm accelerates, the dancers form into two rows, stamp their right feet, bow, and advance with their left feet, making geometric formations. The *kala* dance features a pot symbolizing fecundity. A group of dancers forms a double-tiered circle with other dancers on their shoulders. On top of this tier a man breaks the pot and splashes curds over the naked torsos of the dancers. After this ceremonial opening, the dancers twirl sticks and swords in a feverish battle dance.

Garabā, meaning a votive pot, is the best known religious dance of Gujarāt. It is danced by a group of 50 to 100 women every year for nine nights in honour of the goddess Ambā Mātā, known in other parts of India as Durgā or Kālī. The women move in a circle bending, turning, clapping their hands, and sometimes snapping their fingers. Songs in praise of the goddess accompany this dance.

Of the endless variety of ritualistic folk dances, many have magical significance and are connected with ancient cults. The *karakam* dance of Tamil Nadu state, mainly performed on the annual festival in front of the image of Māriyammai (goddess of pestilence), is to deter her from unleashing an epidemic. Tumbling and leaping, the dancer retains on his head without touching it a pot of uncooked rice surmounted by a tall bamboo frame. People ascribe this feat to the spirit of the deity, which, it is believed, enters his body. The Therayattam festival in Kerala is held to propitiate the gods and demons recognized by the pantheon of the Malayalis. The dancers, arrayed in awe-inspiring costumes and hideous masks, enact weird rituals before the village shrine. A devotee makes an offering of a cock. The dancer grabs it, chops off its head in one stroke, gives a blessing, and hands the bloody gift back to the devotee. This ceremony is punctuated by a prolonged and ponderous dance.

Masked folk dances

The greatest number of masked folk dances are found in Arunachal Pradesh (formerly North East Frontier Agency) union territory of India, where the influence of Tibetan dance may be seen. The yak dance is performed in the Ladākh section of Kashmir and in the southern fringes of the Himalayas near Assam. The dancer impersonating a yak dances with a man mounted on his back. In *sada topo tsen* men wear gorgeous silks, brocades, and long tunics with wide flapping sleeves. Skulls arranged as a diadem are a prominent feature of their grotesquely grinning wooden masks representing spirits of the other world. The dancers rely on powerful, rather slow, twirling movements with hops. The *chhau*, a unique form of masked dance, is preserved by the royal family of the former state of Saraikela in Bihār. The dancer impersonates a god, animal, bird, hunter, rainbow, night, or flower. He acts out a short theme and performs a series of vignettes at the annual Chaitra Parva festival in April. *Chhau* masks have predominantly human features slightly modified to suggest what they are portraying. With serene expressions painted in simple, flat colours, they differ radically from the elaborate facial makeup of *kathākali* or the exaggerated ghoulishness of the Nō and Kandyan masks. His face being expressionless, the *chhau* dancer's body communicates the total emotional and psychological tensions of a character. His feet have a gesture language; his toes are agile, functional, and expressive, like those of an animal. The dancer is mute; no song is sung. Only instrumental music accompanies him. In another form of *chhau*, practiced in the Mayūrbhanj district of Orissa, the actors do not wear masks, but through deliberately stiff and immobile faces they give the illusion of a mask. The style of their dance is vigorous and acrobatic.

Modern Indian dance. While in the West the theatrical elements of spoken words, music, and dance developed independently and evolved in the forms of drama, opera, and ballet, Indian theatrical tradition continued to combine the three in its dramas. Indian films still follow this rule (the heroine suddenly bursts into a song or dances for the hero), which offends Western sensibility, but in fact they are following their own classical and folk tradition. Recently, dance in the form of ballet with complex choreography in the Western sense has emerged as a distinct form.

Modern ballet

Modern Indian ballet started with Uday Shankar, who went to England to study the plastic arts and was chosen by the Russian ballerina Anna Pavlova to be her partner in the ballet *Radha and Krishna*. Young Shankar returned to India fired with enthusiasm. After studying the essentials of the four major styles of classical dance, he created new ballets with complex choreography and music, mixing the sounds from wooden clappers and metal cymbals with those of traditional instruments. He used classical and folk rhythms. Employing Western stage techniques, he presented his ballets with a skill and polish previously unknown to Indian audiences. These ballets included *Shiva-Parvati* and *Lanka Dahan* ("The Burning of Lanka"), in which he used wooden masks from Ceylon. In *Rhythm of Life* (1938) and in *Labour and Machinery* (1939), he employed contemporary political and social themes. He established a culture centre at Almora in 1939 and during its four years' existence created a whole generation of modern dancers.

Shanti Bardhan, a junior colleague of Uday Shankar, produced some of the most imaginative dance-dramas of the modern period. After founding the Little Ballet Troupe in Andheri, Bombay, in 1952 he produced *Ramayana*, in which the actors moved and danced like puppets.

Mohan Khokar

"Ramayana," puppet-style modern dance-drama originally produced and choreographed by Shanti Bardhan, *c.* 1952.

His posthumous production *Panchatantra (The Winning of Friends)* is based on an ancient fable of four friends (Mouse, Turtle, Deer, and Crow), in which he used masks and the mimed movements of animals and birds.

Narendra Sharma and Sachin Shankar, both pupils of Uday Shankar, have continued his tradition. Other important figures who have shaped modern Indian dance include Menaka, Ram Gopal, and Mrinalini Sarabhai, who has experimented with conveying modern themes through the *bhārata-nāṭya* and *kathākali* styles.

Dance-training centres. Dance training in small academies and local *kala kendras* ("art centres") is available all over contemporary India. Most universities have introduced dance as a subject in their curricula. The gurus still impart specialized training to pupils who go to live with them in villages and learn the art over a number of years. But there are many state-run or public-financed training centres organized in the 20th century that attract students from all over the world. Among the most important of these are Kerala Kalamandalam (Kerala Institute of Arts), near Shoranūr; Kalakshetra at Adyar, Tamil Nadu; Kathak Kendra, a dance branch of the Bharatiya Kala Kendra in New Delhi; Triveni Kala Sangam (Centre of Music, Dance, and Painting), at New Delhi; Darpana Academy in Ahmadābād, Gujarāt; Visva-Bharati (founded by Rabindranath Tagore), at Santiniketan, West Bengal;

and the Jawaharlal Nehru Manipuri Dance Academy, at Imphāl.

INDIAN THEATRE

Classical theatre. Classical Sanskrit theatre flourished during the first nine centuries of the Christian era. Aphorisms on acting appear in the writings of Pāṇini, the Sanskrit grammarian of the 5th century BC, and references to actors, dancers, mummers, theatrical companies, and academies are found in Kauṭilya's book on statesmanship, the *Artha-śāstra* (4th century BC). But classical structure, form, and style of acting and production with aesthetic rules were consolidated in Bharata Muni's treatise on dramaturgy, *Nāṭya-śāstra.* Bharata defines drama as a

Bharata Muni's rules

> mimicry of the actions and conduct of people, rich in various emotions, depicting different situations. This relates to actions of men good, bad and indifferent and gives courage, amusement, happiness, and advice to all of them.

Bharata classified drama into ten types. The two most important are *nāṭaka* ("heroic"), which deals with the exalted themes of gods and kings and draws from history or mythology (Kālidāsa's *Śakuntalā* and Bhavabhūti's *Uttararāmacarita* fall into this caterory), and *prakaraṇa* ("social"), in which the dramatist invents a plot dealing with ordinary human beings, such as a courtesan or a woman of low morals (Śūdraka's *Mṛcchakaṭika,* "The Little Clay Cart," belongs to this type). Plays range from one to ten acts. There are many types of one-act plays, including *bhaṇa* ("monologue"), in which a single character carries on a dialogue with an invisible one, and *prahasana* ("farce"), which is classified into two categories: superior and inferior, both dealing with courtesans and crooks. King Mahendravikramavarman's 7th-century-AD *Bhagavad-Ajjukiya* ("The Harlot and the Monk") and *Mattavilāsa* ("Drunken Revelry") are examples of *prahasana.*

Types of classical theatre

There are three structural types of classical theatre: oblong, square, and triangular, each further divided into large, medium, and small sizes. According to the *Nāṭya-śāstra,* the playhouse was "like a mountain cave" with two floors at different levels, small windows so that outside noise and wind would not interfere with the acoustics, and a backstage for actors to do makeup, costumes, and offstage noise effects. Bharata disapproved of a large playhouse and recommended the medium-size structure meant for court productions.

The ancient Hindus insisted on a small playhouse, because dramas were acted in a highly stylized gesture language with subtle movements of eyes and hands. Hindu theatre differed from its Greek counterpart in temperament and method of production. The three unities rigidly followed by the Greeks were totally unknown to Sanskrit dramatists. Less time was consumed by a Greek program of three tragedies and a farce than by a single Sanskrit drama, with its subsidiary plots and wide variety of characters and moods. The Greeks laid emphasis on plot and speech, the Hindus on the four types of acting and visual demonstration. People were audiences to the Greeks and spectators to the Hindus. The aesthetic rules also differed. Aristotle's theory of catharsis bears no resemblance to Bharata's theory of *rasa.* The Greek conception of tragedy is totally absent in Sanskrit dramas, as is the aesthetic principle that prohibits any death or defeat of the hero on stage.

Kinds of Hindu production

There were two types of Hindu productions: the *lokadharmī,* or realistic theatre, with natural presentation of human behaviour and properties catering to the popular taste, and the *nāṭyadharmī,* or stylized drama, which, using gesture language and symbols, was considered more artistic. In *Śakuntalā* the king enters riding an imaginary chariot, and Śakuntalā plucks flowers that are not there; in "The Little Clay Cart" the thief breaks through a nonexistent wall, and Maitreya passes through Vasantasena's seven courtyards by miming.

A classical play traditionally opened with the *nāndī,* a benediction of eight to 12 lines of verse in praise of the gods, after which the *sūtra-dhāra* (stage manager) entered with his wife and described the place and occasion of the action. The last sentence of his prologue served as a bridge leading to the action of the play. In *Śakuntalā* he refers to the bewitching song of his wife, which has made him forget his surroundings as the pursuit of a deer has made the king forget his state affairs. At this point the king enters, riding his hunting chariot, and the spectators are plunged into action of the play.

The *vidūṣaka* (clown) is a noble, good-hearted, blundering fool, the trusted friend of the hero. A bald-headed glutton, comic in speech and manners, he is the darling of the spectators. With the decline of Sanskrit drama the folk theatre in various regional languages inherited the conventions of the opening prayer song, the *sūtra-dhāra,* and the *vidūṣaka.*

The only surviving Sanskrit drama is *kudiyattam,* still performed by the Cakkayars of Kerala. Some principles of the *Nāṭya-śāstra* are evident in their presentations.

The earliest available classical dramas are 13 plays edited in 1912 by Pandit Ganapati Sastri, who dug out their manuscripts in Trivandrum, the capital of Kerala state. These, ascribed to Bhāsa (1st century BC–1st century AD), include the one-act *Ūrubhaṅga* ("The Broken Thigh"), a tragedy that is a departure from Sanskrit convention, and the six-act *Svapnavāsavadatta* ("The Dream of Vāsavadattā").

The most acclaimed dramatist is Kālidāsa. Other important playwrights succeeding him include Harṣa, Mahendravikramavarman, Bhavabhūtī, and Viśakhādatta. An exception is King Śūdraka, whose work is perhaps the most theatrical in the entire Sanskrit range.

"The Little Clay Cart"

The title of "The Little Clay Cart" represents a departure from Sanskrit tradition, in which a *prakaraṇa* was generally named after its hero and heroine. *Mālavikāgnimitra,* for example, is the love story of Princess Mālavikā and King Agnimitra, *Vikramorvaśī* is the tale of King Purūravas and the heavenly nymph Urvaśī, and *Mālatī-Mādhava* is the love drama of Mālatī and Mādhava. Śūdraka, as if to mock tradition, chose an insignificant homely incident—the hero's son playing with a toy cart—and elevated this to the title.

"The Little Clay Cart" has a wide range of characters. The plot does not progress in a straight line but zigzags along a winding path. During its 10 acts the hero does not appear in four of them, the heroine is absent from three, and the lustful villain disappears after the first act until the eighth. Each act is an almost independent play. The device used to link the acts is that of ending them with subtitles that sum up their particular themes or plots.

"The Little Clay Cart" has been successful in the West, whereas Indian audiences, still fed on poetic-flavoured characters and romances of an ethereal type, have favoured *Śakuntalā.* Western audiences find "The Little Clay Cart" more in their own tradition of realism and individualized characterization. Its "lisping villain," gamblers, and rogues have something in common with Shakespeare's comic characters and Molière's crooks. "The Little Clay Cart" is better theatre, whereas *Śakuntalā* is better poetry.

Folk theatre. After the decline of Sanskrit drama, folk theatre developed in various regional languages from the 14th through the 19th centuries. Some conventions and stock characters of classical drama (stage preliminaries, the opening prayer song, the *sūtra-dhāra,* and the *vidūṣaka*) were adopted into folk theatre, which lavishly employs music, dance, drumming, exaggerated makeup, masks, and a singing chorus. Thematically it deals with mythological heroes, medieval romances, and social and political events, and it is a rich store of customs, beliefs, legends, and rituals. It is a "total theatre," invading all the senses of the spectators.

The most crystalized forms are the *jātrā* of Bengal, the *nautanki, rāmlīlā,* and *rāslīlā* of North India, the *bhavai* of Gujarāt, the *tamāshā* of Mahārāshtra, the *terukkūttu* of Tamil Nadu, and the *yakṣagāna* of Kanara.

Places and kinds of folk theatre

Folk theatre is performed in the open on a variety of arena stages; round, square, rectangular, multiple-set. The *bhavai,* enacted on a ground-level circle, and the *jātrā,* on a 16-foot (five-metre) square platform, have gangways that run through the surrounding audience and connect the stage to the dressing room. Actors enter and exit through these gangways, which serve a function similar to the *hanamichi* of the Japanese Kabuki theatre. In the *rāmlīlā,*

the action sometimes occurs simultaneously at various levels on a multiple set. Actors in *nautanki* and *bhavai* sit on the stage in full view instead of exiting and sing or play an instrument as a part of the chorus. In the *rāmlīlā*, the actor playing Rāvaṇa removes his ten-headed mask when he is not acting and continues sitting on his throne, but for the spectators he is theatrically absent. Asides, soliloquies, and monologues abound. Scenes melt into one another, and the action continues in spite of change of locale.

In most folk forms the art of the actor is hereditary. He learns by watching his elders throughout childhood. He starts with drumming, then dancing, plays female roles, and then major roles.

All roles are played by men except that of the *tamāshā* woman, who is always a dancer-singer-actress. Recently, women have started playing female roles in the *jātrā* but have failed to achieve the artistic stature of their professional male counterparts.

In the *rāmlīlā* and *rāslīlā*, the principal characters—Rāma and Krishna—are always played by boys under 14, because tradition decreed they must be pure and innocent.

Mohan Khokar

Rāsīlā folk drama of northern India, watercolour, late 19th century. In the collection of Mohan Khokar, New Delhi. 1.3 × 1.3 m.

They are considered representatives of the gods and are worshipped on these occasions. In the *rāmlīlā* the *vyas* ("director"), present on the stage throughout the performance, prompts and directs the characters loudly enough for the audience to hear. This is not regarded as disturbing because it is an accepted part of the tradition. Adult roles such as Rāvaṇa and Hanuman are sometimes played by the same individual throughout his life.

Non-religious folk dance

Of the nonreligious forms, the *jātrā* and the *tamāshā* are most important. The *jātrā*, also popular in Orissa and eastern Bihār, originated in Bengal in the 15th century as a result of the *bhakti* movement, in which devotees of Krishna went singing and dancing in processions and in their frenzied singing sometimes went into acting trances. This singing with dramatic elements gradually came to be known as *jātrā*, which means "to go in a procession." In the 19th century the *jātrā* became secularized when the repertoire swelled with love stories and social and political themes. Until the beginning of the 20th century, the dialogue was primarily sung. The length has been cut from all night to four hours. The *jātrā* performance consists of action-packed dialogue with only about six songs. The singing chorus is represented by a single character, the *vivek* ("conscience"), who can appear at any moment in the play. He comments on the action, philosophizes, warns of impending dangers, and plays the double of everybody. Through his songs he externalizes the inner feelings of the characters and reveals the inner meaning of their outer actions.

The *tamāshā* (a Persian word meaning "fun," "play," or "spectacle") originated at the beginning of the 18th century in Mahārāshtra as an entertainment for the camping Mughal armies. This theatrical form was created by singing girls and dancers imported from North India and the local acrobats and tumblers of the lower-caste Dombari and Kolhati communities with their traditional manner of singing. It flourished in the courts of Marāthā rulers of the 18th and 19th centuries and attained its artistic apogee during the reign of Bājī Rāo II (1796–1818). Its uninhibited *lavani*-style singing and powerful drumming and dancing give it an erotic flavor. The most famous *tamāshā* poet and performer was Ram Joshi (1762–1812) of Sholāpur, an upper class Brahmin who married the courtesan Bayabai. Another famous singer-poet was Patthe Bapu Rao (1868–1941), a Brahmin who married a beautiful low-caste dancer, Pawala. They were the biggest *tamāshā* stars during the first quarter of the 20th century. The *tamāshā* actress, commonly called the *nautchi* (meaning "nautch girl," or "prostitute") is the life and soul of the performance. Because of their bawdy elements, women never see *tamāshā* plays, nor do respectable men.

In the 20th century, *jātrā* and *tamāshā* both have become highly organized and are commercially run. Troupes are in heavy demand and work for nine months. Over 700 *tamāshā* troupes with 2,000 dancer-actresses tour the rural areas, providing a living for about 40,000 people. The *jātrā* is the most successful commercially. Its star actors draw more than any other professional actor in the theatrical centre of Calcutta.

Puppet theatre of North India

Popular in North India are the *putliwalas* ("puppeteers") of Rājasthān, who operate marionettes made of wood and bright-coloured cloth. The puppet plays deal with kings, lovers, bandits, and princesses of the Mughal period. Generally, the puppeteer and his nephew or son operate the strings from behind, while the puppeteer's wife sits on her haunches in front of the miniature stage playing the drums and commenting on the action. The puppeteer chirps, whimpers, and squeals in animal–bird voices and creates battles and tragic moments, expresses pathos, anger, and laughter. In Andhra Pradesh the puppets, called *tholu bommalata* ("the dance of leather dolls"), are fashioned of translucent, coloured leather. These are projected on a small screen, like colour photographic transparencies. Animals, birds, gods, and demons dominate the screen. The puppeteer manipulates them from behind with two sticks. Strong lamps are arranged so that the size, position, and angle of the puppets change with the distance of the light. They are similar to the *wayang kulit* puppets of Indonesia but are much smaller and quicker moving.

In the absence of a powerful Indian city theatre (with the exception of a few in Calcutta, Bombay, and Tamil Nadu), folk theatre has kept the rural audiences entertained for centuries and has played an important part in the growth of modern theatres in different languages. The 19th-century dramatist Bharatendu Harishchandra, who was responsible for the birth of Hindi drama, used folk conventions—the opening prayer song, tableaux, comic interludes, duets, stylized speech—and combined these with Western theatrical forms in vogue at that time. Parsi companies adapted the popular folk techniques for their extravaganzas and were a major influence until the 1930s. Rabindranath Tagore, rejecting the heavy sets and realistic decor of the commercial companies, created a lyrical theatre of the imagination. Much influenced by the *baul* singers and folk actors of Bengal, he introduced the Singing Bairagi and the Wandering Poet (similar to the *vivek* of the *jātrā*) in his dramas. In the late 20th century, folk theatre has been viewed as a form that can add colour and vitality to contemporary theatre.

Modern theatre. Modern Indian theatre first developed in Bengal at the end of the 18th century as a result of Western influence. The other regional theatres more or less followed Bengal's pattern, and within the next 100 years they took the same meandering path, though they never achieved the same robust growth.

British influence on Bengali theatre

The British conquered Bengal in 1757 and influenced local arts by their educational and political systems. Their clubs performed Shakespeare, Molière, and Restoration comedies, introducing Western dramatic structure and the

proscenium stage to the Indian intelligentsia. With the help of Golak Nath Dass, a local linguist, Gerasim Lebedev, a Russian bandmaster in a British military unit, produced the first Bengali play, *Chhadmabes* ("The Disguise"), in 1795 on a Western-style stage with Bengali players of both sexes. Subsequently, Bengali playwrights began synthesizing Western styles with their own folk and Sanskrit heritage. With growing national consciousness, theatre became a platform for social reform and propaganda against British rule. Among the most important playwrights were Michael Madhu Sudan (1824–73), Dina Bandhu Mitra (1843–87), Girish Chandra Ghosh (1844–1912), and D.L. Roy (1863–1913).

The success of Dina Bandhu Mitra's *Nildarpan* ("Mirror of the Indigo"), dealing with the tyranny of the British indigo planters over the rural Bengali farm labourers, paved the way for professional theatre. The actor-director-writer Girish Chandra Ghosh founded in 1872 the National Theatre, the first Bengali professional company, and took *Nildarpan* on tour, giving performances in the North Indian cities of Delhi and Lucknow. The instigatory speeches and lurid scenes of British brutality resulted in the banning of this production. To overcome censorship difficulties, playwrights turned to historical and mythological themes with veiled symbolism that was clearly understood by Indian audiences. The heroes and villains of these plays came to represent the Indian freedom fighter against the British oppressor. Girish's historical tragedies *Mir Qasim* (1906), *Chhatrapati* (1907), and *Sirajuddaulah* (1909) bring out the tragic grandeur of heroes who fail because of some inner weakness or betrayal of their colleagues. D.L. Roy emphasized the same aspect of nationalism in his historical dramas *Mebarapatan* (The Fall of Mebar), *Shahjahan* (1910), and *Chandragupta* (1911).

Girish introduced professional efficiency and showmanship. His style of acting was flamboyant, with fiery grace. Actors such as Amar Datta and Dani Babu carried his style into the early 1920s. The acting and production methods of the Star, the Minerva, and the Manmohan Theatres (all professional) were modelled on Girish's pioneer work.

The beginning of realism

The first elements of realism were introduced in the 1920s by Sisir Kumar Bhaduri, Naresh Mitra, Ahindra Chowdhuri, and Durga Das Banerji, together with the actresses Probha Devi and Kanka Vati. In his Srirangam Theatre (closed in 1954), Sisir performed two most memorable roles: the again Mughal emperor Aurangzeb and the shrewd Hindu philosopher-politician Cānakya. Sisir's style has been refined by actor-director Sombhu Mitra and his actress wife Tripti, who worked in the Left-wing People's Theatre movement in the 1940s. With other actors they founded the Bahurupee group in 1949 and produced many Tagore plays including *Rakta Karabi* ("Red Oleanders") and *Bisarjan* ("Sacrifice"), so far unattempted by any professional company.

Rabindranath Tagore (1861–1941), steeped in Hindu classics and indigenous folk forms but responsive to European techniques of production, evolved a dramatic form quite different from those of his contemporaries. He directed and acted in his plays along with his cousins, nephews, and students. These productions were staged mostly at his school, Santiniketan, in Bengal as a nonprofessional and experimental theatre. The Calcutta elite and foreign visitors were attracted to these performances.

A painter, musician, actor, and poet, Tagore combined these talents in his productions. He used music and dance as essential elements in his latter years and created the novel opera-dance form in which a chorus sat on the stage and sang while the players acted out their roles in dance and stylized movements. Sometimes Tagore himself sat on a stool acting as the *sūtra-dhāra* and chanted to the accompaniment of music and drum as the dancing players became visual moving pictures.

In northern and western India, theatre developed in the latter half of the 19th century. The Bombay Parsi companies, using Hindi and Urdu, toured all over India. Their spectacular showmanship, based on a dramatic structure of five acts with songs, dances, comic scenes, and declamatory acting, was copied by regional theatres. The Maharashtrian theatre, founded in 1843 by Visnudas Bhave, a singer-composer-wood-carver in the court of the Raja of Sāngli, was developed by powerful dramatists such as Khadilkar and Gadkari, who emphasized Marāthā nationalism. The acting style in Maharashtrian theatre remained melodramatic, passionately arousing audiences to laughter or tears.

In the south, the popularity of dance-dramas has not allowed theatrical realism to flourish. Tamil commercial companies with their song and dance extravaganzas have dominated Andhra Pradesh, Kerala, and Mysore. The most outstanding Tamil company since the independence of India in 1947 has been the T.K.S. Brothers of Madras, famous for trick scenes and gorgeous settings. Also famous is the actor-producer-proprietor Rajamanickam, who specializes in mythological plays with an all-male cast, using horses, chariots, processions, replicas of temples, and even elephants.

Urdu and Hindi drama

Urdu and Hindi drama began with the production of *Indrasabha* by Nawab Wajid Ali Shah in 1855 and was developed by the Parsi theatrical companies until the 1930s.

Parsi theatre was an amalgam of European techniques and local classical forms, folk dramas, farces, and pageants. Mythical titans thundered on the stage. Devils soared in the air, daggers flew, thrones moved, and heroes jumped from high palace walls. Vampire pits, the painted back cloth of a generalized scene, and mechanical devices to operate flying figures were direct copies of the 19th-century Lyceum melodramas and Drury Lane spectacles in London.

The star film actor Prithvi Raj Kapoor founded Prithvi Theatres in Bombay in 1944 and brought robust realism to Hindi drama, then closed down in 1960 with a sense of completion after many tours throughout India. Prithvi's sons, nephews, and old associates worked in his large company, which became a training centre for many actors who later joined the films. Among these was the outstanding stage actress Zohra Sehgal, a former dance partner of Uday Shankar in the 1930s who had tremendous emotional depth and range, rare in actresses on the Hindi stage. Out of Prithvi's eight productions, in which he always played the lead, the best was *Pathan* (1946), which ran for 558 nights. It deals with the friendship between a tribal Muslim khān and a Hindu dewan and is set in the rugged frontier from which Prithvi came. This tragedy of two archtypes in which the khān sacrifices his son to save the life of his friend's son had intensity of action, smoldering passion, and unity of mood and achieved the highest quality of realism on the Hindi stage to this day.

Among the actors who molded regional-language theatres are Shri Narayan Rao Rajhans (popularly known as the Bala Gandharva of the Mahārāshtra stage), Jayashankar Bhojak Sundari of Gujarāt, and Sthanam Narasimhrao of Andhra. All three specialized in female roles and were star attractions during the first quarter of the 20th century.

In the last half of the 20th century, two outstanding actor-directors are Ebrahim Alkazi, director of the National School of Drama in New Delhi, and Utpal Dutt, who founded the Calcutta Little Theatre Group in 1947, which originally performed plays in English and in 1954 changed to productions in Bengali. Dutt is an actor fully committed to the revolutionary ideology of the Chinese Communist leader Mao Tse-tung. He acts on open-air stages in rural areas of Bengal, where he exerts a strong artistic and political influence.

Since Lebedev in 1795 there has been a continuous stream of Western-trained actors and producers who have been revitalizing regional-language theatrical groups. Nawab Wajid Ali Shah had visiting French opera composers in his mid-19th-century court. Tagore did his first opera, *Valmiki Pratibha* ("The Genius of Vālmīkī"), in 1881, after returning from England, where he became familiar with Western harmonies. Prithvi Raj Kapoor, E. Alkazi, and Utpal Dutt all had their earlier training in English productions. Norah Richards, an Irish-born actress who came to the Punjab in 1911, produced in 1914 the first Punjabi play, *Dulhan* ("The Bride"), written by her pupil I.C. Nanda. For 50 years she promoted rural drama and inspired actors and producers, including Prithvi Raj Kapoor.

India's genius still lies in its dance-dramas, which have

a unique form based on centuries of unbroken tradition. There are very few professional theatre companies in the whole of India, but thousands of amateur productions are staged every year by organized groups. Out of this intense experimental activity, the Indians hope a contemporary national theatre will emerge, influenced by Western techniques but distinctly Indian in flavour.

Centres for training

Many centres for theatrical training have been established in the 20th century. Among the most important are the National School of Drama and the Asian Theatre Institute in New Delhi, Sangeet Natak Akademi (National Academy of Music, Dance, and Drama) in New Delhi, and the National Institute for the Performing Arts in Bombay. Bharatiya Natya Sangh, the union of all Indian theatre groups, was founded in 1949 and is centered in New Delhi. Affiliated with UNESCO's branch of the International Theatre Institute, it organizes drama festivals and seminars, as well as serving as a centre for information.

SRI LANKA (CEYLON)

The ritualistic dances of Sri Lanka have attained world fame for their weird mystical beauty. Literary drama has not flourished, because the monks of predominantly Buddhist Sri Lanka shunned theatre. Dramatic activity found expression in exorcism ceremonies and masked dramas that employed mime, song, dance, acrobatics, and bits of prose dialogue. Heavily influenced by India, Sri Lanka's Kandyan dance and *kōlam* plays have South Indian origins. But over the centuries these have been transformed and now have a distinctly Sri Lankan character.

It is difficult to divide Sri Lankan performing arts into dance and drama, because a *kōlam* play uses dance and song, and the devil dance has bits of improvised prose dialogue.

Dance. *Devil dance.* The Buddhists of Sri Lanka still believe in supernatural beings and the healing power of magical rites. Their devil dancing is the expression, however, of pre-Buddhist beliefs.

The devil dance is performed to cure a person gripped by disease, insanity, or bad luck believed to be caused by some malignant spirit and to propitiate demons and deities to bring good fortune. The dancers belong to a lower-caste community and are professional. During their performance the patient lies to one side. Several palm-leaf shrines are constructed outside his house, each dedicated to a particular demon to lure him into the arena. The role of the Vesamuni, king of all demons, is played by the chief exorcist.

Three types of supernatural beings have to be appeased: demons, deities, and others that are half demon, half deity. The most terrible is Riri Yakka (Demon of Blood), who inhabits cremation grounds and graveyards and rides a pig. His belly is smeared with blood, and he has a monkey's face and four clawed hands that hold a parrot, a sword, a rooster, and a human head.

Costumes of the devil dancers

The dancers, all men, each wear around their heads a red cloth fringed with long ribbons of palm leaves hanging down like female hair, a strip of cloth around their chests, 22 yards (20 metres) of thin white cloth wound so skillfully around their hips that it never comes loose during an entire night of violent activity, and clusters of bells fitted around their calves to make a deafening jingle. Their appearance is half female, half male.

The dance is punctuated by little pieces of mimes and magical actions, with naked-chested drummers pounding to the accompaniment of a chorus of singers. The climax is reached when the dancers, holding flaming torches in both hands, whirl and spin, forming circles of fire around themselves. The flames lick their bodies, but they remain unsinged. The dancers leap and dive through the air in seeming defiance of gravity. In this surcharged atmosphere they pause to put on masks representing various demons. These have terrifying expressions. They romp and stomp in circles, describing their identity and purpose of visiting. The particular demon associated with the malady enters the patient's body. The chief exorcist questions, threatens, torutures, beseeches, and offers bribes to appease the demon until it finally leaves and its victim is healed.

The *sanni yakku* dance, exorcising the disease demon, has a series of humorous impersonations. One is of the demon as a beautiful woman, then as a pregnant woman, and finally as a mother. The exorcists ask questions about her pregnancy, and she lists all the respectable men of the village.

Out of many devil-dance ceremonies, most picturesque and important are the *kohomba kankariya* (or "ritual of the god Kohomba"), performed to ensure prosperity and to get rid of pestilence, and the *bali,* danced to propitiate the heavenly beings.

Kandyan dance. The *kandyan* dance, shorn of occult ceremonies, is highly sophisticated and refined. It flourished under the Kandyan kings from the 16th through the 19th centuries, and today it is considered the national dance of Sri Lanka. It has four distinct varieties: The *pantheru, naiyadi, udekki,* and *ves* (the most artistic and renowned). Its energetic movements and postures are reminiscent of India's *kathākali.* Besides the above four styles, there are 18 *vannamas* (dance enactments) including the *gajaga vannama,* depicting the elephant; the *hanuma vannama,* the monkey; and the *mayura vannama,* the peacock. These beautiful animal movements and abstract impersonations have been distilled and perfected over several hundred years.

The *vannamas*

Hindu mythological themes were originally the subjects of *kandyan* dances, the most popular being Rāma's crossing over to Laṅkā with the help of his monkey general and his reunion with Sītā. Gradually, stories of kings and legendary heroes and mimes of birds and forest beasts were introduced. The Kandyan kings elevated the dance to such beauty and skill that the Buddhists began admitting it into their temple courtyards as a tribute to the glory of their religion. It became a part of the annual August Perahera festival, in which a procession of gilded elephants, palanquins, saffron-robed monks, drummers, and chanters move majestically to the Temple of the Tooth in Kandy, where Buddha's tooth is enshrined. The *kandyan* dancers are a glittering attraction as they perform en route to the temple.

Ewing Krainin—Stockpile

Kandyan dancers and drummers from Sri Lanka.

A *kandyan* dancer wears a pagoda-like conical silver headpiece with glistening forehead fringe and huge earpieces, many-stranded bead necklaces of silver and ivory across his naked torso, beaten-silver epaulets on his biceps, and hollow silver anklets filled with silver beads to make them rattle. He spins with sudden leaps and reaches violent climaxes of geometric patterns. The sudden right and left turns of his head make the onlookers dizzy. When telling a story, he sings descriptive passages and enacts them with spurts of dancing.

Masked drama. Out of the four folk-drama forms—*kōlam, sokari, nadagam,* and *pasu*—the most highly de-

veloped and significant is the *kōlam*, in which actors wear brightly painted and intricately carved wooden masks.

The *kōlam*

The word *kōlam* is of Tamil origin and means "costume," "impersonation," or "guise." The performance consists of the masked representation of many isolated characters, such as kings, demons, deities, hunters, animals and birds, the washerman, the police constable, a pregnant woman—a British Museum manuscript concerning the *kōlam* lists 53 such characters. The most terrifying masks are of the demons, with twisted faces, protruding tusks, and cavelike nostrils for snorting fury. The *nāga* demon has a long, flaming, red tongue and dozens of cobras writhing around his head. Some old masks have only one large bulging eye, with a cobra hissing out from one nostril. The design of these masks lies between the strongly stylized tribal masks of Africa and the highly polished, sophisticated masks of the Japanese Nō Theatre. Five basic colours are used—red, blue, yellow, green, and black, the last two for lower rank characters. Exaggerated comicality, distortions, bulges, nightmarish whimsy, bright colours, and artful carving of the masks have been significant factors in keeping this form of drama alive.

The *kōlam* is performed once a year for seven to ten nights, starting at night after dinner and lasting through the early hours of the morning. The performance is generally held in the open courtyard of a house, to the accompaniment of two drummers, an instrumentalist, and a singing chorus with leader. After songs in praise of Lord Buddha and others (including the patron of the show), the *sabhapati* (master of ceremonies) describes the origin of *kōlam*—how an Indian king's pregnant wife expressed a desire to see a masked dance-drama and how a troupe was invited from a distant court. The *sabhapati* then introduces the masked characters as they enter and describes their various vocations and backgrounds.

Out of many, two plays are especially famous: the *Sandakinduru Katava* and the *Gothayimbala Katava*. The former deals with the legendary idyllic love between a half-human, half-bird couple singing and dancing in a forest. The King of Banaras comes hunting and, attracted by the beautiful Kinduri, kills her husband and makes advances to her. Rejected, he is ready to kill her when Lord Buddha appears and brings her husband back to life. In the *Gothayimbala Katava* the beautiful wife of the warrior Gothayimbala bathing in a pond attracts the attention of a demon, who falls in love with her. The enraged husband comes and chops off the demon's head, which, because of its magical power, reunites itself with the body every time it is cut off. Finally, the forest deity comes and rescues the warrior.

The recorded history of *kōlam* is not very old. There is only one known early eye witness account of *kōlam*, that of John Callaway, who in 1829 published 185 verses of a play with a description of the performance and some sketches of the masks and a brief introduction concerning the masquerade. According to Callaway, the dancers did not sing. The chanters described the characters in the third person and sometimes exclaimed to draw the attention of the audience to a particular action. The earliest *kōlam* text is preserved in the Colombo National Museum on palm leaves; another is in the British Museum inscribed on paper. The oldest printed text, edited in 1895 by A.G. Perera, is in the Colombo National Museum Library.

Masks are made of the light woods kaduru (Strychnos nux vomica) and ruk-attana (*Alstonia scholaris*) and after 50 years start decaying; consequently, the earlier masks are no longer in existence.

There has been an important revival of interest in drama in Sri Lanka since independence. E.R. Sarachchandara, a scholar of traditional Ceylonese theatre, has been responsible for a major breakthrough in revitalizing and adapting for the modern stage traditional dramatic forms such as the *kōlam*. Several new playwrights have become prominent in the mid-20th century. Foremost among them is Henry Jayasena. A producer-writer-actor, Jayasena has written and staged plays in Sinhalese and translations of foreign plays. But modern theatre is still very weak.

DANCE AND THEATRE IN KASHMIR

The Vale of Kashmir, predominantly populated by Muslims, has remained aloof from the main cultural currents of India. The ancient caves and temples of Kashmir, however, reveal a strong link with Indian culture at the beginning of the Christian Era. At one time the classical dances of the south are believed to have been practiced. When Islām was introduced, in the 14th century, dancing and theatrical arts were suppressed, being contrary to a strict interpretation of the Qur'ān. These arts survived only in folk forms and were performed principally at marriage ceremonies. The popular *hafiza* dance performed by Kashmiri women at weddings and festivals to the accompaniment of *sufiana kalam* (devotional music of the Muslim mystics known as Ṣūfīs) was banned in the 1920s by the ruling maharaja, who felt this dance was becoming too sensual. It was replaced by the *bacha nagma*, performed by young boys dressed like women. A popular entertainment at parties and festivals, it is also customarily included in modern stage plays.

In contrast to its natural scenic richness, Kashmir is theatrically a pauper. Theatrical productions are generally amateurish, since there is no regular performing company or any tradition of civic theatre.

There is only the *bhand jashna* ("festival of clowns"), a 300- to 400-year-old genre of Kashmiri folk theatre. Performed in village squares, it satirizes social situations through dance, music and clowning.

Bhand jashna

The Kashmiri-language theatre was founded in 1947, when a new national consciousness, the aftermath of the independence of the Indian subcontinent from Britain, inspired playwrights and folk actors to dramatize topical events and create a "visual newspaper" for the people. Left-wing propaganda plays such as *Zamin Sanz* ("Who Owns the Land?") and *Jangbaaz* ("The Warmonger"), though mediocre, had topical interest. Notable among those who tried their hand at writing for the stage is the poet Nadim, author of two operas, *Bambur-yambarzal (The Bumblebee)* and *Himal Nagraj (The Beautiful Woman and the Snake Prince)*.

Since the 1960s, the Jammu and Kashmir Academy of Art, Culture, and Languages has been struggling to promote theatre in the Kashmiri and Dogri languages, but with little success. Its emphasis is on literary dramas and folk-dance festivals of regional appeal.

PAKISTAN

Muslim culture has frowned upon the performing arts, with the result that there is no Arabic or Persian classical theatre. The only possible sources of drama were the Persian passion plays dealing with the martyrdom of Ḥuysayn (grandson of Muḥammad) in the desert of Karbalā' in 680 AD, which have inspired some Urdu playwrights. Pakistan, a Muslim country, therefore either had to find a theatrical heritage in Urdu and Bengali theatre, which had been flourishing in India long before the partition, or look to the West. It did both. The Urdu-language theatre of Pakistan had started in the Lucknow court of Nawab Wajid Ali Shah in 1855 and was nurtured by both Muslim and Hindu artists. In Pakistan the *kathak* style is preferred because of its strong Muslim flavour and Mughal court associations. Cut off from Hinduism and its lore, Pakistani performers use these Indian classical dance styles to interpret the aspirations of a young nation, while their folk dances express the character of Pakistan's rural culture.

Folk dance. Pakistan's dances are virile and explosive. *Bhangra* and Khaṭak are the most popular. Khaṭak is a dance of the tribal Pathans, known for their hospitality and feuds in the rugged hills of the northwest. It originated in zealous preparations for raids and celebrations of victories. In the 20th century, any joyous event is the occasion for this community dance. The Pathans, dressed in baggy *salwars*, embroidered waistcoats, and skullcapped turbans, perform it holding a rifle in both hands. In a frenzy they spin and somersault, float and whirl, with sudden bursts of swordplay to the accompaniment of drums and pipes. Because of its popularity, Khaṭak is presented to visiting dignitaries and for this purpose has been refined into choreographed productions.

Khaṭak

Bhangra, folk dance of the Punjab region of Pakistan and India.
Lustig—FPG

Female dances

Important dances by women are the *sammi, kikli, giddha,* and *luddi.* Except for the *sammi,* which has a slow rhythm accompanied by a sad song because of its association with the tragic love legend of Princess Sammi and Prince Dhola, all the other forms are charged with energy and fast rhythms. The *kikli* is performed by teen-age girls in groups of two. The partners cross their arms, interlock their fingers, and, touching the toes of their feet, stretch backward and whirl. The *giddha* is danced in a circle, the participants keeping the rhythm by clapping their hands. Two women impulsively leave the circle, jump into the centre, and perform a hilarious mimetic dance enacting a *boli* (two-line song) and again join the circle to dance in a ring and allow another couple to take the centre. In the *luddi,* women click their fingers and clap their hands, moving in a circle by jumps and half-turns and accelerating their rhythm by stamping their feet.

Performing arts in the Punjab. The genius of Punjabis finds expression in love stories, lusty dancing, and humour. The *mirasis* (professional wits), *naqalias* (mummers), and *domanis* (female singer-actresses) are professional performers belonging to the lower classes. They exploit all the tricks of exaggeration, absurdity, malapropism, comic gags, and lewd references. In the performance of a *naqal* (comic sketch), two people constitute a troupe. The leader holds a leather folder and slaps his foolish partner, who leads his master to a hilarious situation through absurd replies. Expert in mime and clowning, these character types are distantly related to the Western court fool and the commedia dell'arte.

Theatre in Pakistan. Urdu theatre grew out of a spectacular production of *Indrasabha* ("The Heavenly Court of Indra"), an operatic drama written by the poet Agha Hasan Amanat and produced in 1855 in the palace courtyard of the last Nawab of Oudh, Wajid Ali Shah. The story deals with the love of a fairy and Prince Gulfam. The fairy takes her lover to heaven where the angry and jealous Indra hurls him down to earth. Finally, the fairy, through her songs and dances, wins the heart of Indra, and the two lovers are united. Wajid Ali Shah, an expert *kathak* dancer and author of many valuable treatises on stage techniques, composed some of the melodies and dances for his production and used folk conventions, gorgeous costumes, elaborate settings, and gold-inlaid masks. *Indrasabha* was a fantastic success; it was translated into almost all the regional languages, with many local variants. Its characters—Sabaz Pari (Green Fairy), Kala Deo (Black Devil), Lal Deo (Red Devil)—became a part of the theatrical vocabulary of the subcontinent.

Parsi theatre. During the second half of the 19th century, Urdu was the main spoken and written language of the northern half of the subcontinent and understood in almost all the principal cities. The Parsis (originally Zoroastrians from Iran who settled on the coast of Bombay), comprising a wealthy community with sharp business acumen, were the pioneers in establishing a commercial theatre, that lasted from 1873 to 1935 and influenced all the other regional theatres. Though located mainly in Bombay and Calcutta, the Parsi companies toured the subcontinent with huge staffs, sets, and an army of players.

The best known playwright of this period is Agha Hashr (1876–1935), a poet-dramatist of flamboyant imagination and superb craftsmanship. Among his famous plays are *Sita Banbas,* based on an incident from the *Rāmāyaṇa; Bilwa Mangal,* a social play on the life of a poet, whose blind passion for a prostitute results in remorse; and *Aankh ka Nasha* ("The Witchery of the Eyes"), about the treachery of a prostitute's love, with realistic dialogue of a brothel. Many of Hashr's plays were adapted from Shakespeare: *Sufayd Khūn* ("White Blood") was modelled on *King Lear,* and *Khūn-e Nāḥaq* ("The Innocent Murder") on *Hamlet.* His last play, *Rustam-o-Sohrab,* the tragic story of two legendary Persian heroes, Rustam and his son Sohrab, is a drama of passion and fatal irony.

Parsi productions

Productions by Parsi theatrical companies were large-budgeted affairs. Plays opened with the actors in full makeup and costume, their hands folded and eyes closed, singing a prayer song in praise of some deity, and generally ended in a tableau. Sometimes at curtain call the director rearranged the tableau in a split second and offered a variant. Actors were required to know singing, dancing, music, acrobatics, and fencing and to possess strong voices and good physical bearing. In improvised auditoriums with bad acoustics and packed with more than 2,000 people, actors' voices reached the farthest spectator. Plays began at 10 o'clock and lasted until dawn, moving from comedy to tragedy, from pathos to farce, from songs to the rattle of swords, all interspersed with moral lessons and rhyming epigrams. The droll humour and realism of the comic interludes remain unsurpassed in contemporary Urdu drama. Important playwrights of this period were Narain Prasad Betab, Mian Zarif, and Munshi Mohammed Dil of Lucknow. All took inspiration from Hindu mythology and Persian legends, transforming these tales into powerful dramas.

Imtiaz Ali Taj (1900–70) was a bridge between Agha Hashr and contemporary Pakistani playwrights. His *Anarkali* (1922), the tragic love story of a harem girl, Anarkali, and Crown Prince Salim (son of Akbar the Great), unfolds the love-hate relationship of a domineering emperor and his rebellious son. Brilliant in treatment and character analysis, this play has been staged hundreds of times by amateur groups and has entered the list of Urdu classics.

In the absence of a professional company, Urdu theatre has found it difficult to strike roots. After 1947 many Muslim actors and writers were absorbed by the Indian film industry in Bombay, and they found it difficult to adjust their great talent to amateur theatrical clubs. All the same, plays have been staged in Karāchi, Lahore, and Rāwalpindi. The best productions have been those dealing with topical themes—refugee problems, new adjustments, the corrupt bureaucracy, the Kashmir issue, and other sociopolitical issues. Agha Babar in Rāwalpindi produced *Burra Sahib* (1961: "The Big Boss"), an adaptation of Gogol's *Government Inspector,* setting it in Pakistan. *Tere Kuce se Jub Hum Nikle* ("Thrown Out of Your Lane"), by Naseer Shamshi, describes the pathetic condition of an aristocratic family in Delhi that is forced to leave home because of communal riots. In *Lal Qile se Lalukhet Tak* ("From the Red Fort to Lalukhet"), by Khwajah Moinuddin, the comedy arises out of the pitiable condition of the refugees who leave their well-settled existence in Delhi dreaming of prosperity, take a tedious journey, and arrive homeless in Karāchi to find shelter in thatched hovels. Ali Ahmed, an avant-garde actor-director in Karāchi, presents his plays with polished stagecraft and esoteric appeal.

Lahore remains the centre of amateur theatre based on the tradition of the late directors A.S. Bokhari and G.D.

Sondhi, both former principals of the Government College in Lahore. In 1942 G.D. Sondhi built the Open-Air Theatre, situated on a small artificial hillock in the Lawrence Gardens and perhaps the best in all of South Asia. It has remained the centre of dramatic contests and festivals and is a favourite of visiting dancers and actors.

The actor-playwright Rafi Peer, with his knowledge of Western theatre as a result of his training in Berlin in the 1930s, has helped to develop Pakistani theatre. Professional in approach, he has produced radio and stage plays and has been a critical colleague of A.S. Bokhari and Imtiaz in the revival of amateur theatre.

Radio and television plays. Plays are being written for radio and television that are readily adaptable for the stage, and vice versa. Saadat Hussan Mantoo (1912–55), the greatest writer of short stories and author of over 100 radio plays and features, is still the model for contemporary writers for plot construction, bitter realism, and whimsical dialogue. His collection of plays (1942–45), including *Mantoo ke Dramay* ("Mantoo's Plays"), *Ao* ("Come"), and *Teen Aurten* ("Three Women"), have flashes of the then-unborn Theatre of the Absurd.

More dramas are written in Urdu today than are staged. The turnover is large because of their generally amateurish character and short runs. There is no major professional centre for the training of actors, nor a school for stagecraft and production. Notwithstanding, the young directors and playwrights have been enthusiastic about establishing a permanent Urdu stage.

BANGLADESH

East Bengal continued the folk *jātrā* and used this form for themes concerning current political problems and historical events. A successful example of the latter is *Nawab Sirajuddaulah,* which deals with the fall of the last Muslim ruler of Bengal in 1757 through betrayal by his ambitious brother-in-law Mīr Jaʿfar, who joined the British. This *jātrā* is popular with both rural and urban audiences. Tales of Muslim kings and lovers from Persian legends also have been rendered into *jātrās.*

Contemporary theatre inherits the tradition of the prepartition Bengali stage. The poet-playwright Nazrul Islam followed the tradition of Tagore's verse plays and dance operas. Inspired by left-wing ideology, he wrote for the People's Theatre in East Bengal, championing the cause of the poor farmer. He has dealt with psychological problems and inner tensions in his *Shilpi* ("The Artist"), in which the artist is torn between love for his wife and for his art. Especially popular are historical themes of political significance, inspiring Muslims who for centuries were subjugated by the Hindus of East Bengal. Ebrahim Khan wrote *Kamal Pasha* (1926), a play about the Turkish liberator, a symbol of hope and reawakening, and *Anwar Pasha,* about the downfall of Anwar (Enver), who could not cope with the new historical forces.

Bangladesh has a solid acting tradition and a rich repertoire of Bengali plays. Its amateur stage has professional actors, and it retains the impassioned lyricism and power of the mainstream of Bengali tradition. (B.Ga.)

Visual arts

VISUAL ARTS OF THE INDIAN SUBCONTINENT

Indian art is the term commonly used to designate the art of the Indian subcontinent, which includes the present political divisions of India, Kashmir, Pakistan, and Bangladesh. Although a relationship between political history and the history of Indian art before the advent of Islām is at best problematical, a brief review will provide a broad context. The earliest urban culture of the subcontinent is represented by the Indus Valley civilization (*c.* 2500–1800 BC), which possessed several flourishing cities not only in the Indus Valley but also in Gujarāt and Rājasthān. The circumstances in which this culture came to an end are obscure. Although there is no clear proof of historical continuity, scholars have noticed several striking similarities between this early culture and features of later Indian civilization. The period immediately following the urban Indus Valley civilization is marked by a variety of essentially rural cultures. A second urbanization began to occur only around the 6th century BC, when flourishing cities started to reappear, particularly in the Gangetic Basin. The Buddha lived and preached in this period, which culminated in the great Maurya Empire, whose relatively few works are the earliest surviving remnants of monumental art. The Maurya dynast Aśoka (died 238 BC) is considered the greatest of Buddhist kings; and the majority of the monuments of the next 500 years appear to be dedicated to the Buddhist faith, though iconographical and other details suggest that the art also drew heavily on popular religion.

The Maurya Empire spread over almost all of what is modern India and Pakistan. Territories as extensive were never possessed by any other dynasty. With its fall, the empire broke up into a number of states ruled by many dynasties, some of which acquired considerable power and fame for varying periods of time. Among these, the Śuṅgas (*c.* 2nd–1st century BC) in the north and the longer-lived Sātavāhanas in the Deccan and the south are particularly noteworthy. Though these kings were Hindu by religion, Buddhist monuments form the great majority of surviving works.

Toward the end of the 1st century BC, northern India was subjected to a series of invasions by Scythian tribes, resulting finally in the establishment of the vast Kushān (Kuṣāṇa) empire, of which Mathurā was an important centre. The new rulers seemed to have followed Indian faiths, the great emperor Kaniṣka (*c.* AD 78) being a devout Buddhist. The schools of Gandhāra and Mathurā flourished during their rule, and, though much of the work is dedicated to the Buddhist religion, the foundations of later Hindu iconography were also laid in this period. While the Kushān dynasty was sovereign in the north, the Sātavāhanas continued to rule in the south. The bulk of the work at Amarāvatī was produced during their hegemony.

The Golden Age of India

Around the mid-4th century, the Gupta dynasty, of indigenous origin, rapidly expanded its power, uprooting the last remnants of foreign rule and succeeding in bringing almost all of northern India under its sway. In the Deccan there arose at the same time the equally powerful Vākāṭakas, with whom the Guptas appear to have had friendly relations. The period extending from the 4th through the 5th centuries is marked by the most flourishing artistic activities. In addition to the Buddhist monuments, there are the first strong indications of specifically Hindu patronage. Works of remarkable beauty and elegance were produced in this period, which is commonly called the Golden Age of India.

The disintegration of these two empires toward the close of the 5th and the 6th centuries ushered in what has been called the medieval period (*c.* 8th–12th centuries), marked by the appearance of a large number of states and dynasties, often at war with each other. Their rise to power and their decline was part of a constantly recurring process, for none of them was able to hold onto a position of even relative paramountcy for any extended period of time. In the north, the great dynasties were the Gurjara-Pratīhāras, whose empire at its greatest equalled that of the Guptas; the Pālas, who ruled chiefly over northeastern India; and various other dynasties, such as the Kalacuris, the Candelas, and the Paramāras of north central India, the Cāhamānas of Rājasthān, the Cālukyas of Gujarāt. In the Deccan, also, several dynasties rose and fell, the most powerful of which were the Cālukyas of Bādāmi, the Rāṣṭrakūṭas, and the Cālukyas of Kalyāṇī. They were often at war not only with their powerful neighbours to the north but also with the great Pallava and Cōḷa kingdoms of southern India. Most of the dynasties of medieval India were Hindu, though some Jaina and a very few Buddhist kings are also known. The various faiths, however, existed in comparative harmony; and Buddhist and Jaina monuments continued to be built, though most of the surviving works are Hindu.

Although the effects of constant struggle were not as devastating as one might expect, largely as a result of the institutionalization of war and its confinement to appropriate castes, the Hindu kingdoms fell easy prey to the Islāmic invasions, which began as early as the 8th century

AD but gathered strength only in the 11th century. By the end of the 12th century, almost all of northern India had been conquered. Islāmic advances in the south were checked for a while by the Vijayanagara dynasty, but with its collapse almost all of India fell under various degrees of Islāmic hegemony. Large Hindu kingdoms enjoying differing degrees of independence continued to exist chiefly in Rājasthān and portions of southern India, but overall political supremacy was vested with the Islāmic states. The Muslim powers were also divided into many kingdoms, despite attempts made by the sultanate of Delhi, and later by the Mughals, to achieve paramountcy over large portions of India. These attempts were successful only for short periods of time. Although the initial impact of Islām on Indian art was generally destructive, Islāmic influences entering India were gradually transformed in the new environment and eventually resulted in the flowering of an extremely rich and important aspect of the Indian genius.

Ascendency of the European powers

The ascendency of the European powers in the 18th century, culminating in the establishment of the British Empire, laid the foundation of modern India's contacts with the West. As a whole, the European advent was marked by a relative insensitivity to native art traditions, but rising nationalism attempted a conscious revival of Indian art toward the end of the 19th century. In modern times, the absorption of European influence is a more natural, freer process that affects artistic development in a vital and profound way.

General characteristics. *The unity of Indian art.* Indian art is spread over a subcontinent and has a long, very productive history; but it nevertheless shows a remarkable unity and consistency. Works produced in the several geographical and cultural regions possess decidedly individual characteristics but at the same time have sufficient elements in common to justify their being considered manifestations of a general style. The existence of this style is evidence of the essential cultural unity of the subcontinent and to the uninterrupted contact between the various geographical units, at least from the historical period onward. Developments in one area have been quickly reflected in the others. The regional idioms have contributed to the richness of Indian art, and the mutual influences exercised by them have been responsible for the multi-faceted development of that art throughout the course of its long life.

The style of Indian art is largely determined not by a dynasty but by conditions of time and space. It has, essentially, a geographical rather than a dynastic basis, which is to say that the evolution of regional schools appears to have been largely independent of any particular dynasty that happened to rule over a specific region. The style does not change because of the conquest of one area by another dynasty; rather the influences exercised by one area on another are usually through the agency of factors other than conquest. Instances in which dynastic patronage changed the nature of a style are very few and confined mostly to the Islāmic period. The political history of India is itself quite vague, and the areas in possession of a dynasty at various points in its history are even less susceptible to precise definition. For all these reasons, the classification of Indian art adopted here is not based on dynasties, for such a division has little meaning. Nevertheless, names of certain dynasties are used, for these have passed into common usage. When this is done, however, the name must be understood as little more than a convenient way of labelling a particular period.

The materials of Indian art. Indian art employs various materials, such as wood, brick, clay, stone, and metal. Most wooden monuments of the early period have perished but have been imitated in stone. Clay and brick were also abundantly used; but, particularly in later times, the favoured material seems to have been stone, in the dressing (facing and smoothing) and carving of which the Indian artist attained great excellence. The material may have influenced the form somewhat, but essentially Indian art tends to impose the form on the material. Thus, materials are generally regarded as interchangeable: wooden and clay forms are imitated in stone, and stone is imitated in bronze, and in turn stone sculpture assumes qualities appropriate to metal. It is as though the nature of the material presented a challenge that had to be met and overcome. At the same time, Indian art stresses the plasticity of forms; sculpture is generally characterized by emphatic mass and volume; architecture is often sculpture on a colossal scale; and the elements of painting, particularly of the early period, are modelled by line and colour.

Emphasis on the plasticity of forms

Indian and foreign art. Thanks to its geographical situation, the Indian subcontinent has been constantly fed by artistic traditions emanating from West and Central Asia. The Indian artist has shown a remarkable capacity for accepting these foreign influences naturally and assimilating and transforming them to accord with the nature of his own style. The process occurred frequently: in the Maurya period; in the two centuries after Christ, when the Kushān dynasty attained imperial supremacy in the north; and at a much later period, in the 16th century, when the Mughals patronized a new school of architecture and painting.

Just as India received influences, so it transmitted its own art abroad, particularly to Ceylon and the countries of Southeast Asia. Developments of great importance were therby precipitated in Ceylon, Burma, Thailand, Indonesia, and Indochina, where the reinterpretation of Indian influences resulted in the creation of works of great originality.

Indian art and religion. Indian art is religious inasmuch as it is largely dedicated to the service of one of several great religions. It may be didactic or edificatory as is the relief sculpture of the two centuries before and after Christ; or, by representing the divinity in symbolic form (whether architectural or figural), its purpose may be to induce contemplation and thereby put the worshipper in communication with the divine. Not all Indian art, however, is purely religious, and some of it is only nominally so. There were periods when humanistic currents flowed strongly under the guise of edificatory or contemplative imagery, the art inspired by and delighting in the life of this world.

Although Indian art is religious, there is no such thing as a sectarian Hindu or Buddhist art, for style is a function of time and place and not of religion. Thus it is not strictly correct to speak of Hindu or Buddhist art, but, rather, of Indian art that happens to render Hindu or Buddhist themes. For example, an image of Vishnu and an image of Buddha of the same period are stylistically the same, religion having little to do with the mode of artistic expression. Nor should this be surprising in view of the fact that the artists belonged to nondenominational guilds, ready to lend their services to any patron, whether Hindu, Buddhist, or Jaina.

Religion, symbolism, and abstraction

The religious nature of Indian art accounts to some extent for its essentially symbolic and abstract nature. It scrupulously avoids illusionistic effects, evoked by imitation of the physical and ephemeral world of the senses; instead, objects are made in imitation of ideal, divine prototypes, whose source is the inner world of the mind. This attitude may account for the relative absence of portraiture and for the fact that, even when it is attempted, the emphasis is on the ideal person behind the human lineaments rather than on the physical likeness.

The artist and patron. Works of art in India were produced by artists at the behest of a patron, who might commission an object to worship for spiritual or material ends, in fulfillment of a vow, for the discharge of virtues enjoined by scripture, or even for personal glory. Once the artist received his commission, he fashioned the work of art according to his skill, gained by apprenticeship, and the written canons of his art, which possessed a holy character. There were prescribed rules for proportionate measurement, iconography, and the like, often with a symbolic significance. This is not to say that the individual artist was invariably aware of the symbolic meaning of the prescribed standards, based as these were on complex metaphysical and theological considerations; but the symbolism nevertheless formed part of the fabric of his work, ready to add an extra dimension of meaning to the initiated and knowledgeable spectator.

In these conditions it is not surprising that the artist as

Early Indian architecture and sculpture

Stūpa III at Sānchi, Madhya Pradesh, sandstone, 1st century BC.

Detail of a sandstone relief sculpture from the *torana* of Stūpa I at Sānchi, Madhya Pradesh, 1st century BC.

Fragment of a soapstone disc from Kausāmbī, Uttar Pradesh, 3rd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh. Height 5.7 cm.

Abduction scene, terra-cotta relief from Kausāmbī, Uttar Pradesh, *c.* 2nd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh.

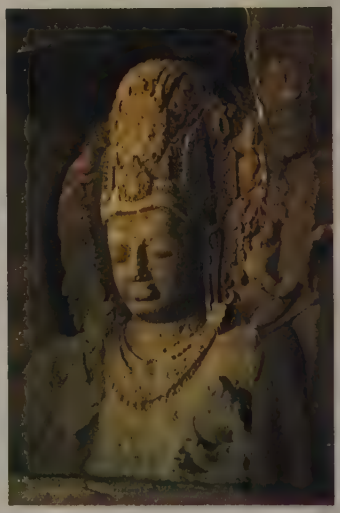
Plates 1–8 P. Chandra

Viṣṇu lying on the serpent Sesa; sandstone relief panel on the Viṣṇu temple at Deogarh, Uttar Pradesh, 5th century AD.

Indian art of the Golden Age: 5th to 6th centuries AD

(Below) The Pārvatī temple at Nācnā Kuthārā, Madhya Pradesh, sandstone, 5th century AD. (Left) Amorous couples and lotus scroll, detail of the doorframe of the Pārvatī temple.

Lotus scroll painted on the ceiling of Cave 2 at Ajanta, Mahārāshtra, 5th century AD.

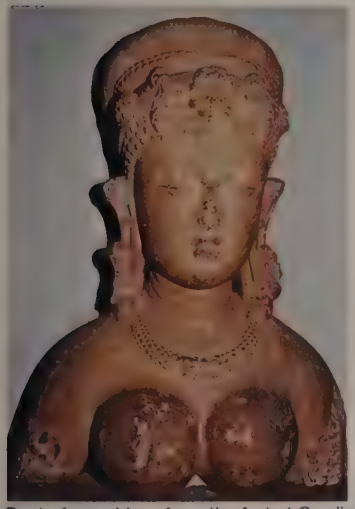


Śiva-Mahesamūrti, the main image in the cave Temple at Elephanta, Mahārāshtra, 6th century AD.

Medieval sculpture of North and South India

The hermitage by the Ganges, detail of a granite relief depicting the penance of Arjuna, from Mahābalipuram, Tamil Nadu, early 7th century AD.

The god Śiva, detail of a doorframe from Singhpur, Madhya Pradesh, sandstone, 10th century AD.



Bust of a goddess from the fort at Gwalior, Madhya Pradesh, *c.* 9th century AD. In the Gwalior Museum, Madhya Pradesh. Height 54 cm.

Śiva and his consort, bronze sculpture from Tiruvenkadu, Tamil Nadu, early 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height of male 106 cm., height of female 94 cm.

Seated Avalokiteśvara, gilt bronze sculpture from Nālandā, Bihār, 8th century AD. In the Nālandā Museum, Bihār. Height 28 cm.

Medieval temple architecture: styles of North and South India

Temple with *kutina* superstructure, the Tālapurīśvara at Panamalai, Tamil Nadu, early 8th century AD.

Temples, tank, and *gopura* of the Śiva temple at Cidambaram, Tamil Nadu, 12th–13th century AD.

Group of temples at Badoli, Rājasthān, 10th century AD.

Temple with *bhūmija* superstructure, the Udayaeśvara (or Nīlakaṇṭhhaeśvara) at Udayapur, Madhya Pradesh, *c.* AD 1059–82.

Temple with *latina* superstructure at Umrī, Madhya Pradesh, 9th century AD.

(Above) The Citragupta temple at Khajurāho, Madhya Pradesh, sandstone, 11th century AD. (Left) Detail of the temple wall.

(Left) The Brhadīsvara temple at Thanjāvūr, Tamil Nadu, early 11th century AD. (Right) Detail of the temple wall.

The Keśava temple at Somnāthpur, Karnataka, *c.* AD 1268.

The monk Kālaka addressing the Sāhi king, detail from a folio from a *Kālakācāryakathā* manuscript, Western Indian style, late 13th century AD. In the Prince of Wales Museum of Western India, Bombay. Dimensions of miniature 7.6 × 7.6 cm.

Ladies in conversation, detail from a folio from a *Mahābhārata* manuscript, AD 1516. In the collection of the Asiatic Society of Bombay. Dimensions of miniature 10.2 × 10.2 cm.

A prince and his lady, Rajasthani style, Mālwa, mid-17th century AD. In a private collection.

Krsna and Rādhā, miniature from a series illustrating the *Gītagovinda*, Rajasthani style, Mewār, mid-17th century AD. In a private collection.

A hill chief smoking, Pahari style, Basohlī, late 17th century AD. In the National Museum of India, New Delhi.

The Mughal style and its influences

A queen on a hunting expedition, Pahari style, Kāngra, *c.* AD 1760. In the National Museum of India, New Delhi.

Court scene by Basāvan, folio from an illustrated manuscript of the *Anwār-e Suhayli*, Mughal style, AD 1596–97. In the Bharat Kala Bhavan, Vārānasi, Uttar Pradesh. 24.8 × 13.9 cm.

The feast of Nauroz at Jahāngīr's court, Mughal style, *c.* AD 1615. In the collection of the Reza Library, Rāmpur, Uttar Pradesh.

The musical mode Megha-Malāra, Rajasthani style, Būndi, late 17th century AD. In a private collection.

Copper bowl with hunting scene, Mughal, *c.* AD 1583. In the Prince of Wales Museum of Western India, Bombay.

The so-called Dīvān-e Khāss at Fatehpur Sīkri, red sandstone, *c.* AD 1585.

Mughal architecture and decorative arts

The tomb of Humāyūn at Delhi, red sandstone and marble, *c.* AD 1564.

Detail of a silk tapestry, Mughal, early 17th century AD. In the Prince of Wales Museum of Western India, Bombay.

Jade dish inlaid with semiprecious stones and gold, Mughal, mid-17th century AD. In the Prince of Wales Museum of Western India, Bombay.

a person is for the most part anonymous, very few names of artists having survived. It was the skill with which the work of art was made to conform to established ideals, rather than the artist who possessed the skill, that held the place of first importance.

The appreciation of Indian art. According to Indian aesthetic theory, a work of art possesses distinct "flavours" (*rasa*), the "tasting" of which constitutes the aesthetic experience. Because the work of art operates at various levels, granting to the spectator what he is capable of receiving by virtue of his intellectual and emotional preparation, the appreciation of the beauty of form and line is considered an appropriate activity of the educated and cultured man. The supreme aesthetic experience, however, is believed to be much deeper and cognate to the experience of the Godhead. From this point of view, the work of art is in a sense irrelevant and unnecessary for a person at a high level of spiritual progress; and for the devout layman its excellence is measured by its efficacy in promoting spiritual development.

Architecture. The favoured material of early Indian architecture appears to have been wood, but little has survived the rigours of the climate. Wooden forms, however, affected work in other mediums and were sometimes quite literally copied, as, for example, in early cave temples of western India. The principles of wooden construction also played an important part in determining the shape of Indian architecture and its various elements and components.

Effect of wooden forms on other mediums

Baked or sun-dried brick has a history as ancient as that of wood; among the earliest remains are buildings excavated at sites of the Indus Valley civilization. The use of brick is once again evident from about the 6th century BC, and its popularity was undiminished in subsequent centuries. Many brick monuments have been discovered, particularly in areas in which good clay was easily available, such as the Gangetic Basin. Although more durable than wood, few brick buildings from before the 5th century AD have survived in a good state of preservation.

Traditions of stone architecture appear to be more recent than wood or brick, the earliest examples of the use of dressed stone for building purposes not predating the 6th century BC. The Indian architect, however, soon gained great proficiency in its use, and, by the 7th century AD, the use of stone for monumental buildings of considerable size had become quite popular. The preference for stone can also be seen in Islāmic monuments of India, which contrast markedly with the brick and tile structures popular in neighbouring West Asia.

Most surviving examples of Indian architecture before the Islāmic period are of a religious nature, consisting mainly of Buddhist shrines, or *stūpa*s, and temples. Monastic residences give some idea of civil architecture, but, surprisingly, very few examples of palaces and secular dwellings have been found.

Indus Valley civilization (c. 2500–1800 BC). From excavated remains, it is clear that the Indus Valley civilization possessed a flourishing urban architecture. The major cities associated with the civilization, notably Mohenjodaro, Harappā, and Kalibangan, were laid out on a grid pattern and had provisions for an advanced drainage system. The residential buildings, which were serviceable enough, were mainly brick and consisted of an open patio flanked by rooms. For monumental architecture, the evidence is slight, the most important being a "sacred" tank (thought to be for ritual ablution) and associated structures. Corbel vaulting (arches supported by brackets projecting from the wall) was known, and, to a limited extent, timber was used together with brick; whatever architectural ornamentation existed must have been of brick or plaster.

The Maurya period (c. 321–185 BC). The state of Indian architecture in the period between the Indus Valley civilization and the rise of the Maurya Empire is largely unknown since most work was done in such perishable material as wood or brick. Excavations at Rājgīr, Kauśāmbī, and other sites, however, testify to the existence of fortified cities with *stūpa*s, monasteries, and temples of the type found at the later Maurya sites of Nagarī and Vidiśā; and there is evidence of the use of dressed stone in a palace excavated at Kauśāmbī. Considering the power of the Maurya Empire and the extensive territory it controlled, the architectural remains are remarkably few. The most important are *stūpa*s (later enlarged) such as a famous example of Sānchi; the ruins of a hall excavated at the site of Kumrāhar in Patna (ancient Pāṭaliputra), the capital city; and a series of rock-cut caves in the Barābar and Nāgārjunī hills near Gayā, which are interesting because they preserve in the more permanent rock some types of wooden buildings popular at that time.

The *stūpa*

The *stūpa*, the most typical monument of the Buddhist faith, consists essentially of a domical mound in which sacred relics are enshrined. Its origins are traced to mounds, or tumuli, raised over the buried remains of the dead that were found in India even before the rise of Buddhism: *Stūpa*s appear to have had a regular architectural form in the Maurya period: the mound was sometimes provided with a parasol surrounded by a miniature railing on the top, raised on a terrace, and the whole surrounded by a large railing consisting of posts, crossbars, and a coping (the capping on the top course), all secured by tenons and mortices in a technique appropriate to craftsmanship in wood. The essential feature of the *stūpa*, however, always remained the domical mound, the other elements being optional.

Along with *stūpa*s were erected roofless, or hypaethral, shrines enclosing a sacred object such as a tree or an altar. Temples of brick and timber with vaulted or domical roofs were also constructed, on plans that were generally elliptical, circular, quadrilateral, or apsidal (*i.e.*, having an apse, or semicircular plan, at the sanctum end). These structures have not survived, but some idea of their shape has been obtained from the excavated foundations and the few examples imitating wooden originals that were cut into the rock, notably the Sudāmā and the Lomas Ṛṣi caves in the Nāgārjunī and Barābar hills near Gayā. The latter has an intersesting entrance showing an edged barrel-vault roof (an arch shaped like a half cylinder) in profile supported on raked pillars, the ogee arch (an arch with curving sides, concave above and convex toward the top) so formed filled with a trellis to let in light and air. The interiors of most caves are highly polished and consist of two chambers: a shrine, elliptical or circular in plan with a domed roof (Sudāmā cave); and an adjacent antechamber, roughly rectangular and provided with a barrel vault. Remains of structural buildings have been excavated at Bairāt and Vidiśā, where wood and brick shrines with timber domes and vaults once existed. A temple (No. 40) at Sānchi was apsidal in plan and perhaps had a barrel-vault roof of timber.

A hall excavated at Kumrāhar in Patna had a high wooden platform of most excellent workmanship, on which stood eight rows of 10 columns each, which once supported a second story. Only one stone pillar has been recovered, and it is circular in shape and made of sandstone that has been polished to a high lustre. The capitals that topped them must have been similar to others found in neighbouring Lohanipur and almost certainly consisted of one or two pairs of addorsed (set back to back) animals, recalling Persepolitan examples. Indeed, there is much about Maurya architecture and sculpture to suggest Iranian influence, however substantially transformed in the Indian environment.

Early Indian architecture (2nd century BC–3rd century AD). Except for *stūpa*s, architectural remains from the 2nd century BC (downfall of the Maurya dynasty) to the 4th century AD (rise of the Gupta dynasty) continue to be rare, indicating that most of the work was done in brick and timber. Once again, examples cut into the rock and closely imitating wooden forms give a fairly accurate idea of at least some types of buildings in this period.

The *stūpa*s become progressively larger and more elaborate. The railings continue to imitate wooden construction and are often profusely carved, as at Bhārhut, Sānchi II, and Amarāvatī. These were also provided with elaborate gateways, consisting of posts supporting from one to three architraves, again imitating wooden forms and covered with sculpture (Bhārahut, Sānchi I, III). In the course of

time an attempt was made to give height to the *stūpas* by multiplying the terraces that supported the dome and by increasing the number of parasols on top. In Gandhāra and southeastern India, particularly, sculptured decoration was extended to the *stūpa* proper, so that terraces, drums, and domes—as well as railing—were decorated with figural and ornamental sculpture in bas-relief. *Stūpas* in Gandhāra were not provided with railings but, instead, had rows of small temples arranged on a rectangular plan.

West Indian cave temples

Cave temples of western India, cut into the scarp of the Western Ghāts and stretching from Gujarāt to southern Mahārāshtra, constitute the most extensive architectural remains of the period. Two main types of buildings can be distinguished, the temple proper (*caitya*) and the monastery (*vihāra, saṅghārāma*). The former is generally an apsidal hall with a central nave flanked by aisles. The apse is covered by a half dome; and two rows of pillars, which demarcate the nave, support a barrel-vault roof that covers the rest of the building. In the apsidal end is placed the object to be worshipped, generally a *stūpa*, the hall being meant for the gathered congregation. In front of the hall is a porch, separated from it by a screen wall provided with a door of considerable size, together with an arched opening on top clearly derived from wooden buildings of the Lomas Ṛṣi type and permitting air and dim light to filter into the interior. Other influences of wooden construction are equally striking, particularly in the vaulting ribs that cover the entire ceiling and that are sometimes actually of wood, as at Bhājā, where the pillars are also raked in imitation of the exigencies of wooden construction. The pillars are generally octagonal with a pot-shaped base and a capital of addorsed animals placed on a bell-shaped, or campaniform, lotus in the Maurya tradition. The most significent example is at Kārli, dating approximately to the closing years of the 1st century BC. The Bhājā *caitya* is certainly the earliest, and important examples are to be found at Beḍsā, Kondane, Pītalkhorā, Ajantā, and Nāsik. Toward the end of the period, a quadrilateral plan appears more and more frequently, as, for example, at Kuda and Sailarwāḍī.

In addition to the *caitya*, or temple proper, numerous monasteries (*vihāras*) are also cut into the rock. These are generally provided with a pillared porch and a screen wall pierced with doorways leading into the interior, which consists of a "courtyard" or congregation hall in the three walls of which are the monks' cells. The surviving rock-cut examples are all of one story, though the facade of the great monastery at Pitalkhorā simulates a building of several stories.

Monasteries carved into the rock are also known from Orissa (Udayagiri-Khandagiri), in eastern India. These are much humbler than their counterparts in western India, and consist of a row of cells that open out into a porch, the hall being absent. At Uparkot in Junāgadh, Gujarāt, is a remarkable series of rock-cut structures dating from the 3rd–4th century AD, which appear to be secular in character and in all probability served as royal pleasure houses.

The large number of representations of buildings found on relief sculpture from sites such as Bhārhut, Sānchi, Mathurā, and Amarāvatī are a rich source of information about early Indian architecture. They depict walled and moated cities with massive gates, elaborate multi-storied residences, pavilions with a variety of domes, together with the simple, thatched-roofed huts that remained the basis of most Indian architectural forms. A striking feature of this early Indian architecture is the consistent and profuse use of arched windows and doors, which are extremely important elements of the architectural decor.

The Gupta period (4th–6th centuries AD). Dating toward the close of the 4th and the beginning of the 5th century AD is a series of temples that marks the opening phase of an architecture that is no longer content with merely imitating wooden building but initiates a new movement, ultimately leading to the great and elaborate temples of the 8th century onward.

Two main temple types of the Gupta period

Two main temple types have been distinguished in the Gupta period. The first consists of a square, dark sanctum with a small, pillared porch in front, both covered with flat roofs. This type of temple answers the simplest needs of worship, a chamber to house the deity and a roof to shelter the devotee. Temple No. 17 at Sānchi is a classic example of this flat-roofed type. The plain walls are of ashlar masonry (made up of squared stone blocks), composed of sizable blocks, which are spanned by large slabs that constitute the ceiling. The pillars of the porch have a campaniform lotus capital, one of the last times this form appears in Indian architecture. Another temple of this type is the Kaṅkālī Devī shrine at Tigowā, which has more elaborate pillars, provided with the overflowing vase, or the vase-and-foliage (*ghaṭa-pallava*), capital that became the basic north Indian order.

It is the second type of temple that points the way to future developments. It also has a square sanctum, or cella, but instead of a flat roof there is a pyramidal superstructure (*śikhara*). Among the most interesting examples are a brick temple at Bhītargaon and the Vishnu temple at Deogarh, built entirely of stone. The pyramidal superstructure of each consists essentially of piled-up cornice moldings of diminishing size, which are decorated primarily with *candraśālā* (ogee arch) ornament derived from the arched windows and doors so frequently found in the centuries immediately before and after Christ. The sanctums of both temples are square in plan, with three sides provided with central offsets (vertical buttress-like projections) that extend from the base of the walls right up to the top of the *śikhara* (spire); the section of the central offset that extends across the wall is conceived in the form of a niche, in which is placed an image. The Deogarh temple is also noteworthy for the large terrace with four corner shrines (now ruined) on which it is placed, prefiguring the quincunx, or *pañcāyatana*, grouping (one structure in each corner and one in the middle) popular in the later period. The doorway surround, too, is very

P. Chandra

Carved sandstone doorframe of the Vishnu temple at Deogarh, Uttar Pradesh, India, 5th century AD.

elaborate, carved with several bands carrying floral and figural motifs. At the base of the surround are rows of worshippers, and in the crossette (projection at the corner) on top are images of graceful river goddesses.

The Pārvatī Devī temple at Nācnā Kuṭhārā, also of this period, is interesting for the covered circumambulatory provided around the sanctum and the large hall in front. When first discovered, the temple had an entire chamber above the sanctum (which subsequently collapsed). Though provided with a door, there seems to have been no access to it; thus, for all practical purposes it constituted a false story and, aside from a symbolic meaning,

served no other purpose than to emphasize the importance of the sanctum. The principle of gaining height not by the superimposition of ornamental cornice moldings with *candraśālā* decoration but by the multiplication of stories, each imitating the story below, also distinguished the later architectural style of southern India.

The great Mahābodhi temple at Buddh Gaya, commemorating the spot where the Buddha attained enlightenment, though burdened with later restorations, is essentially a temple of this period. It has a particularly majestic *śikhara*, decorated with ornamental niches and *candraśālās*, rising over a square sanctum to a great height.

Along with temples, *stūpas* continued to be built. These also aspired to height, which was achieved by multiplication and heightening of the supporting terraces and elongating the drum and dome. A good example of this new form is the Dhamekh *stūpa* at Sārnāth. Along more conventional lines, but quite elaborate, are the brick *stūpas* in Sind, notably a fine example at Mirpur Khās.

The rock-cut temple and monastery tradition also continued in this period, notably in western India, where the excavations—especially at Ajantā—acquire extreme richness and magnificence. The monasteries are characterized by the introduction of images into some of the cells, so that they partake of the nature of temples instead of being simple residences. Temples with an apsidal plan and barrel-vault roofs, however, soon went out of fashion, and are found very rarely in the subsequent period. The early 5th-century cave temples at Udayagiri, Madhya Pradesh, are similar to the simple flat-roofed temples with a hall and are not descended from ancient traditions as preserved in western India.

Period of the great stone temples

Medieval temple architecture. Architectural styles initiated during the 5th and 6th centuries found their fullest expression in the medieval period (particularly from the 9th to the 11th centuries), when great stone temples were built. Two main types can be broadly distinguished, one found generally in northern India, the other in southern India. To these can be added a third type, sharing features of both and found in Karnataka and the Deccan. These three types have been identified by some scholars with the *nāgara, drāviḍa,* and *vesara* classes referred to in some Sanskrit texts, though the actual meaning of these terms is far from clear. In spite of the havoc wrought by the destructive Islāmic invasions, particularly in the Indo-Gangetic Plains, an extremely large number of monuments have survived in almost every other part of India, particularly in the south, and these continue to be discovered and recorded to the present day.

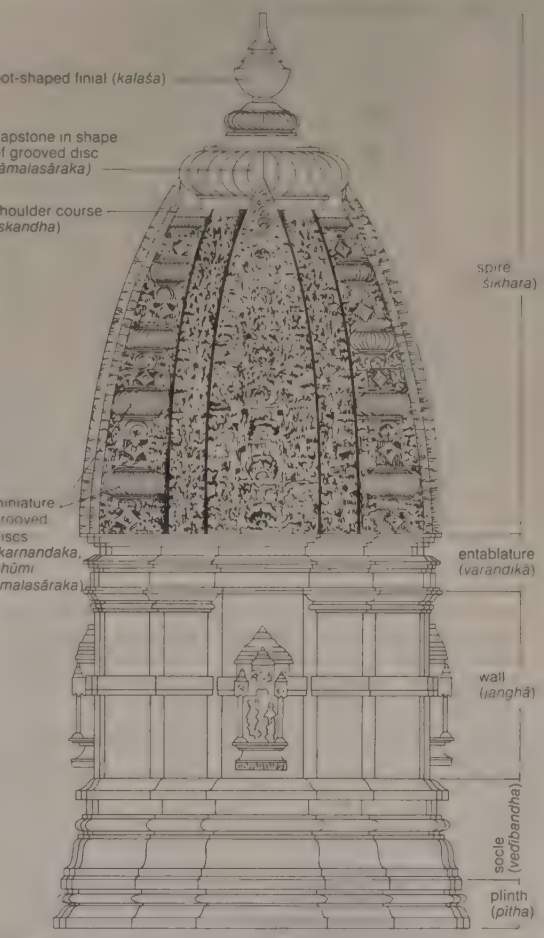
Medieval temple architecture: North Indian style. North Indian temples generally consist of a sanctum enshrining the main image, usually square in plan and shaped like a hollow cube, and one or more halls (called *maṇḍapas*), aligned along a horizontal axis. The sanctum may or may not have an ambulatory, but it is invariably dark, the only opening being the entrance door. The doorway surrounds are richly decorated with bands of figural, floral, and geometrical ornament and with river-goddess groups at the base. A vestibule (*antarāla*) connects the sanctum to the halls, which are of two broad types: the *gūḍhamaṇḍapas*, which are enclosed by walls, light and air let in through windows or doors; and open halls, which are provided with balustrades rather than walls and are consequently lighter and airier. The sanctum almost invariably, and the *maṇḍapas* generally, have *śikharas*; those on the sanctum, appropriately, are the most dominant in any grouping. Internally, the sanctum has a flat ceiling; the *śikhara* is solid theoretically, though hollow chambers to which there is no access must be left within its body to lessen the weight. The ceilings of the halls, supported by carved pillars, are coffered (decorated with sunken panels) and of extremely rich design.

The sanctum is often set on a raised base, or a plinth (*pīṭha*), above which is a foundation block, or socle (*vedībandha*), decorated with a distinct series of moldings; above the *vedībandha* rise the walls proper (*jaṅghā*), which are capped by a cornice or a series of cornice moldings (*varaṇḍikā*), above which rises the *śikhara.* One, three, and sometimes more projections extend all the way from the base of the temple up the walls to the top of the *śikhara.* The central offset (*bhadra*) is the largest and generally carries an image in a niche; the other projections (*rathas*), too, are often decorated with statuary.

The entire temple complex, including sanctum, halls, and attendant shrines, may be raised on a terrace (*jagatī*), which is sometimes of considerable height and size. The attendant shrines—generally four—are placed at the corners of the terrace, forming a *pañcāyatana,* or quincunx, arrangement that is fairly widespread. The temple complex may be surrounded by a wall with an arched doorway (*toraṇa*).

Śikhara types

The *śikhara* is the most distinctive part of the North Indian temple and provides the basis for the most useful and instructive classification. The two basic types are called *latina* and *phāmsanā.* Curvilinear in outline, the *latina*



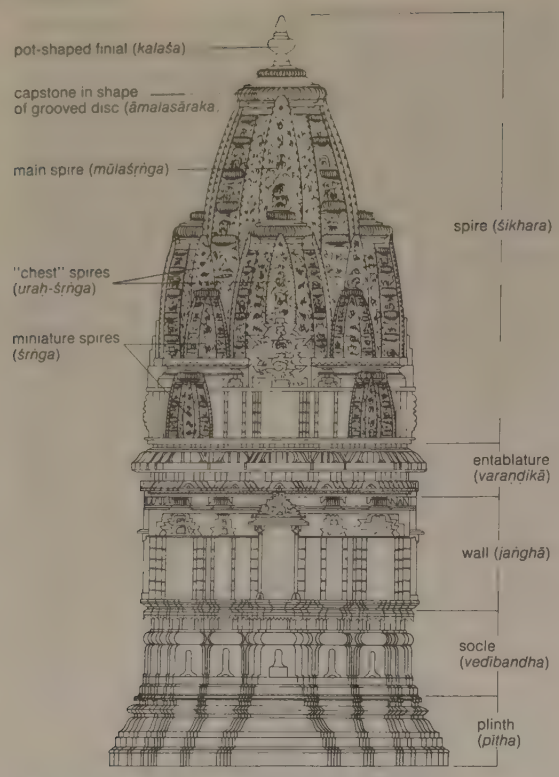
Elevation of a North Indian temple with the *latina* type of superstructure.

is composed of a series of superimposed horizontal roof slabs and has offsets called *latās.* The edges of the *śikhara* are interrupted at intervals with grooved discs, each one demarcating a "story." The surface of the entire *śikhara* is covered with a creeper-like tracery, or interlaced work, composed of diminutive ornamental *candraśālās.*

The *śikhara* is truncated at the top and capped by a shoulder course (*skandha*), above which is a circular necking (*grīvā*), carrying a large grooved disc called the *āmalasāraka.* On it rests a pot and a crowning finial (*kalaśa*).

Unlike the *latina,* the *phāmsanā śikhara* is rectilinear rather than curvilinear in outine, and it is lower in height. It is composed of horizontal slabs, like the *latina,* but is capped by a bell-shaped member called the *ghaṇṭā.* The surface of this type of *śikhara* may have projections, like the *latina śikhara,* and be adorned with a variety of architectural ornament.

From the 10th century onward, the *śekharī* type of spire, an elaboration of the *latina* type, became increasingly popular. In its developed form it consisted of a



Elevation of a North Indian temple with the *śekharī* type of superstructure.

central *latina* spire (*mūlaśṛṅga*) with one or more rows of half spires added on the sides (*uraḥ-śṛṅga*) and the base strung with miniature spires (*śṛṅgas*). The corners, too, are sometimes filled with quarter spires, the whole mass of carved masonry recalling a mountain with a cluster of subsidiary peaks.

The *latina* and *śekharī* spires are generally found on the sanctum, while the *phāmsanā* and its variants are usually confined to the *maṇḍapas*, or halls. The sanctum spires also have a large and prominent projection in front (*śukanāsā*), generally rising above the vestibule (*antarāla*). These projections are essentially large ogee arches of complex form, which often contain the image of the presiding deity.

A particularly rich and pleasing variety of North Indian *śikhara*, popular in Mālwa, western India, and northern Deccan, is the *bhūmija* type. It has a central projection on each of the four faces, the quadrants so formed filled with miniature spires in vertical and horizontal rows right up to the top.

The *bhūmija śikhara*

Although basically reflecting a homogeneous architectural style, temple architecture in northern India developed a number of distinct regional schools. A detailed elucidation of all has yet to be made, but among the most important are the styles of Orissa, central India, Rājasthān, and Gujarāt. The style of Kashmir is distinct from the rest of northern India in several respects, and hardly any examples of the great schools that flourished in modern Uttar Pradesh, Bihār, and Bengal are left standing. The North Indian style also extended for some time into the Karnataka (formerly Karṇāṭa) territory, situated in the southern Deccan, though the architecture of Tamil Nadu was relatively unaffected by it.

Medieval temple architecture: North Indian style of Orissa. The greatest centre of this school is the ancient city of Bhuvaneśvara, in which are concentrated almost 100 examples of the style, both great and small, ranging in date from the 7th to the 13th century. Among the earliest is the Paraśurāmeśvara temple (7th–8th century), with a heavy, stately *latina śikhara*, to which is attached a rectangular *gūḍhamaṇḍapa* with double sloping roofs. The walls are richly carved, but the interiors, as in almost all examples of the style, are left plain. The Mukteśvara temple (10th century), which has a hall with a *phāmsanā* roof, is the product of the most exquisite workmanship. The enclosing wall and the arched entrance, or *toraṇa*, are still present, giving a clear idea of a temple with all its parts fully preserved. The Brahmeśvara temple, which is dated on the basis of an inscription to the mid-10th century, is a *pañcāyatana*, with subsidiary shrines at all of the corners. The most magnificent building, however, is the great Liṅgarāja temple (11th century), an achievement of Orissan architecture in full flower. The *latina* spire soars to a considerable height (over 125 feet [40 metres]); the wall is divided into two horizontal rows, or registers, replete with statuary; and the

P Chandra

Mukteśvara temple at Bhubaneswar, Orissa, India, late 10th century AD.

attached hall is exquisitely and minutely carved. The most famous of all Orissan temples, however, is the colossal building at Konārak, dedicated to Sūrya, the sun god. The temple and its accompanying hall are conceived in the form of a great chariot drawn by horses. The *śikhara* over the sanctum has entirely collapsed; and all that survives are the ruins of the sanctum and the *gūḍhamaṇḍapa*, or enclosed hall, and also a separate dancing hall. Of these, the *gūḍhamaṇḍapa* is now the most conspicuous, its gigantic *phāmsanā śikhara* rising in three stages and adorned with colossal figures of musicians and dancers.

Because the Orissan style usually favours a *latina śikhara* over the sanctum, the *śekharī* spire of the Rājrānī temple (11th century) at Bhuvaneśvara (Bhubaneswar) is quite exceptional. Of particular interest as a late survival of early building traditions is the Vaitāl Deul (8th century), the sanctum of which is rectangular in plan, its *śikhara* imitating a pointed barrel vault. Besides Bhuvaneśvara, important groups of temples are to be found at Khiching and Mukhalingam.

Medieval temple architecture: North Indian style of central India. The area roughly covered by the modern state of Madhya Pradesh was the centre of several vigorous schools of architecture, of which at least four have been identified. The first flourished at Gwalior and adjacent areas (ancient Gopādrī); the second in modern Bundelkhand, known in ancient times as Jejākabhukti; the third in the eastern and southeastern parts in the ancient country of Ḍāhala, of which Tripurī, near modern Jabalpur, was the capital; and the fourth in the west, in an area bordering Gujarāt and Rājasthān in the fertile land of Mālava (Mālwa).

The Gwalior temples

The earliest examples in the Gwalior area are a group of small shrines at Naresar, a few miles from Gwalior proper; dating to the 8th century, the shrines have *latina* spires and sparsely ornamented walls. In the 9th century a series of magnificent temples was built, including the Mālā-de at Gyāraspur, the Śiva temples at Mahuā and Indore, and a temple dedicated to an unidentified mother goddess at Barwa-Sāgar. The period appears to have been one of experimentation, a variety of plans and spires having been tried. The Mālā-de temple is an early example of the *śekharī* type in its formative stages; the Indore temple has a star-shaped plan; and the Barwa-Sāgar example has a twin *latina* spire over a rectangular sanctum. The masonry work is of the finest quality and the architectural ornament is crisply carved. (The figural sculptures are few.) The temple at Umrī, with a *latina* spire, is small and exquisitely finished; but the largest and perhaps the finest temple is the Telī-kā-Mandir on Gwalior Fort, rectangular in plan and capped by a pointed barrel vault, recalling once again the survival of ancient roof forms. The walls are decorated with niches (empty at present) topped by tall pediments (triangular gable ornament).

The style of this region became increasingly elaborate from the 10th century, during the supremacy of the Kacchapaghāṭa dynasty. The many examples from this period are distinguished by a low plinth and rich sculptural decoration on the walls. Outstanding among them are the Kākan-maḍh at Suhāniā (1015–35) and the Sās-Bahū temple (completed 1093) in Gwalior Fort. The several temples at Surwāyā and Kadwāhā, though smaller in size, are distinguished for their extremely rich and elegant workmanship.

The style is best represented by a large group of temples at Khajurāho, the capital of the Candella dynasty, though examples are also to be found in Mahoba and at several other sites in the Jhānsi district of Uttar Pradesh, notably Chāndpur and Dudhai. All of the distinctive characteristics of the fully developed style can be seen in the Lakṣmaṇa temple at Khajurāho (dated 941), which is a *pañcāyatana* placed on a tall terrace enclosed by walls. The sanctum has an ambulatory and, facing it, a series of halls, including the *gūḍhamaṇḍapa*, a porch, and a small intermediate hall. Both the ambulatory and the *gūḍhamaṇḍapa* are provided with lateral, balconied arms, or transepts, which let in light and air. Each hall has its own pyramidal *śikhara*, all skillfully correlated to ascend gradually to the main *śekharī* spire over the sanctum. Extraordinary richness of carving, both in the interior and on the exterior,

The Lakṣmaṇa temple

Lakṣmaṇa temple at Khajurāho, Madhya Pradesh, India, *c.* AD 941.
P Chandra

where the walls carry as many as three rows of sculpture, and a skillful handling of the main spire to suggest ascent are distinguishing features of the style. The largest temple of the group, very similar to the Lakṣmaṇa, is the Kandāriyā Mahādeo; and among the most distinguished are the Viśvanātha and the Pārśvanātha temples. The Dūlādeo temple, which does not have an ambulatory, represents the closing phase of the group and probably belongs to the 12th century.

The earliest temples of the Ḍāhala area, dating from the 8th–9th century, are the simple shrines at Bāndhogarh, which consist of a sanctum with *latina* spire and porch. To the 10th century, when the local Kalacuri dynasty was rapidly gaining power, belong the remarkable Śiva temples at Chandrehe and Masaun, the former being circular in plan, with a *latina* spire covered with rich *candraśālā* tracery. The Golā Math at Maihar has the more conventional square sanctum, with a very elegant *latina śikhara*, the walls of which are adorned with two rows of figural sculpture. There must have existed at Gurgī a large number of temples, though all of them now are in total ruin. Judging from a colossal image of Śiva-Pārvatī and a huge entrance, which have somehow survived, the main temple must have been of very great size. Another important site is Amarkantak, where there are a large group of temples, the most important of which is the Karṇa. Although generally of the 11th century, they are quite simple, lacking the rich sculptural decoration so characteristic of the period. By contrast, the Virāṭeśvara temple at Sohāgpur, with an unusually tall and narrow *śekharī* spire, is covered with sculptural ornamentation as rich as that of Khajurāho.

The Mālava region, ruled largely by the Paramāra dynasty, appears to have been the first to develop the *bhūmija* type of *śikhara* (10th century). The finest and most representative group of these structures is at Un. Though, unfortunately, they are considerably damaged, judging from the remains, they must have been very elegant struc-

Temples of the Mālava area

tures. The best preserved and easily the finest *bhūmija* temple is the Udayeśvara (1059–82), situated at Udaipur in Madhya Pradesh. The *śikhara*, based on a stellate plan, is divided into quadrants by four *latās*, or offsets, each one of which has five rows of aediculae. The large hall has three entrance porches, one to the front and two to the sides, and walls that are richly carved. The whole complex, including seven subsidiary shrines, is placed on a broad, tall platform. The Siddheśvara temple at Nemāwar (early 12th century) is even larger than the Udayeśvara, though the proportions are not as well balanced and the quality of the carving is inferior. Structures in the *bhūmija* manner continued to be made in Mālava up to the 15th century; the Malvai temple at Alīrājpur is a good example of the late phase.

From Mālava, the *bhūmija* style spread to the neighbouring regions. To the north in Rājasthān, the Mahānāleśvara temple at Menāl (*c.* 11th century), the Sun temple at Jhālrapātan (11th century), the Śiva temple at Rāmgarh (12th century), and the Uṇḍeśvara temple (12th century) at Bījoliān are important examples. To the west, in Gujarāt, are temples at Limkheda and Sarnāl of the 11th and 12th centuries. The style was particularly favoured in Mahārāshtra, to the south. Among surviving examples, the most impressive is the Ambarnāth temple near Bombay (11th century); Balsāṇe and Sinnar also have pleasing temples. The style continued up to the 16th century, many examples having been found in north Deccan and Berār. The *bhūmija* style also spread to the east of Mālava; the Bhāṇḍ Dewal at Arang (11th century), for example, is a Ḍāhala adaptation.

Medieval temple architecture: North Indian style of Rājasthān. A group of temples at Osiān, dating to about the 8th century, represents adequately the opening phases of medieval temple architecture in Rājasthān. They stand on high terraces and consist of a sanctum, a hall, and a porch. The sanctum is generally square and has a *latina* spire. The walls, with one central and two subsidiary projections, are decorated with sculpture, often placed in niches with tall pediments. The halls are generally of the open variety, provided with balustrades rather than walls, so that the interiors are well lit. The surrounds of the doorway sanctum are quite elaborate, with four or five bands of decoration and the usual river-goddess groups at the base. The pillars, with *ghaṭa-pallava* (vase-and-foliage) capitals, are also decorated, richness of sculpture and architectural elaboration being a characteristic of this group of monuments. The Mahāvīra temple, which is the largest, belongs to the 8th century, though renovated in later times, when the *toraṇa* (gateway) and the *śikhara* were added. Other important temples are Harihara Nos. 1, 2, and 3 and two temples dedicated to Vishnu. The ruined Harshat Mātā temple at Ābānerī, of a slightly later date (*c.* 800), was erected on three stepped terraces of great size and is remarkable for the exquisite quality of the carving. Some of the finest temples of the style date from the 10th century, the most important of which are the Ghaṭeśvara temple at Bāḍolī and the Ambikā Mātā temple at Jagat. The simple but beautiful Bāḍolī temple consists of a sanctum with a *latina* superstructure and an open hall with six pillars and two pilasters (columns that project a third of their width or less from the wall) supporting a *phāmsanā* spire. Only the central projections of the sanctum walls are decorated with niches containing sculpture. A large open hall was built in front of the temple at a later date. The Ambikā-Mātā temple at Jagat, of the mid-10th century, is exceptionally fine. It consists of a sanctum, a *gūḍhamaṇḍapa*, or enclosed hall, and a parapeted porch with projecting eaves. The walls of the sanctum and the hall are covered with fine sculpture, the superstructures being of the *śekharī* and the *phāmsanā* types.

Growing resemblance between styles of Rājasthān and Gujarāt

Temples, too numerous to mention, dating from the 10th and—to an even greater extent—the 11th century onward, are found throughout Rājasthān. The styles of Rājasthān and neighbouring Gujarāt during these centuries grew closer and closer together until the differences between them were gradually obliterated. This coalescence resulted in the emergence of a composite style found throughout Gujarāt and Rājasthān. Temples situated in the two areas are discussed separately here, but this is for the sake of convenience and does not signify any real stylistic difference.

The temples at Kirāḍu in Rājasthān, dating from the late 10th and 11th centuries, are early examples of the style shared by Rājasthān and Gujarāt. The Someśvara temple (*c.* 1020) is the most important and clearly shows the movement toward increasing elaboration and ornamentation. Each of the constituent parts became more complex; the moldings of the plinth, for example, are multiplied to include bands of elephants, horses, and soldiers. The walls are covered with sculpture, and the spire is of the rich *śekharī* type. Situated in Rājasthān, but again in the composite style, are the extraordinarily sumptuous temples known as the Vimala Vasahī (1031) and the Lūṇa Vasahī (1230) at Mt. Ābū. The Vimala Vasahī consists of a sanctum, a *gūḍhamaṇḍapa*, and a magnificent assembly hall added in mid-12th century. The plain, uncarved exterior walls of the rectangular enclosure of the temple have on the inside rows of cells containing images of divinities. The interiors are very richly carved, the coffered ceilings loaded with a wealth of detail. The Lūṇa Vasahī is even more elaborate, though the quality of the work had begun to decline perceptibly.

Traditional architecture continued even after the Islāmic invasions, particularly during the reign of Rāṇā Kumbhā of Mewār (*c.* 1430–69). During this period, the tall nine-storied Kīrttistambha and other temples at Chitor and also the great Chaumukha temple at Ranakpur (1438) were built.

Medieval temple architecture: North Indian style of Gujarāt. Gujarāt was the home of one of the richest regional styles of northern India. A temple at Gop (*c.* 600), with a tall terrace and a cylindrical sanctum with high walls capped by a *phāmsanā* roof, and other temples in Saurāshṭra show the formative phases of the style. Its distinctive features are clear in an interesting group of temples from Roḍā (*c.* 8th century). The sanctum is square in plan and has *latina* spires that are weighty and majestic. The walls are relatively plain, with niches, housing images, provided only on the central projection. The masonry work is exceptionally good, a characteristic of Gujarāt architecture throughout its history. The Rāṇakdevī temple at Wadhwān, of the early 10th century, is also characterized by plain walls and a *latina* spire, while the Śiva temple at Kerākoṭ has a *śekharī* spire and also a *gūḍhamaṇḍapa*. The great Sun Temple at Modhera, datable to the early years of the 11th century, represents a fully developed Gujarāt style of great magnificence. The temple consists of a sanctum (now in ruins), a *gūḍhamaṇḍapa*, an open hall of extraordinary richness, and an arched entrance in front of which was the great tank. The Navalakhā temple at Sejakpur continued this tradition. The Rudramāla at Siddhapur, the most magnificent temple of the 12th century, is now in a much ruined condition, with only the *toraṇa* (gateway) and some subsidiary structures remaining. Successively damaged and rebuilt, the Somanātha at Prabhāsa Patan was the most famous temple of Gujarāt, its best known structure dating from the time of Kumārapāla (mid-12th century). It has been now dismantled, but a great temple built at the site in recent years testifies to the survival of ancient traditions in modern Gujarāt.

The temple at Gop

The hills of Satrunjaya and Girnār house veritable temple cities. Most of the shrines, which are of late date, are picturesque but otherwise of little significance. With the Islāmic conquest, the Gujarāt architect adapted his considerable skills to meet the needs of a patron of different religion and quickly produced a totally successful Indian version of Islāmic architecture.

Medieval temple architecture: North Indian style of Karnataka. The North Indian style was largely confined to India above the Vindhyas, though for a short period it also flourished in a region of southern India known as Karnataka from ancient times and now largely part of Karnataka (formerly Mysore) state. Here, temples of the northern and the southern styles are found next to each other, notably at Aihole and Pattadkal. The earliest appears to be the Lāḍh Khān at Aihole, closely related

to the 5th-century temple at Nāchnā Kutharā in northern India. The northern style was also cultivated at Pattadkal, where the most important examples are the Kāśīviśvanātha, the Galaganātha, and the Pāpanātha. Ālampur, now in Andhra Pradesh, has eight temples of the northern style with *latina* spires. These belong to the late 7th and early 8th centuries and are the finest and among the last examples of the northern style in southern India.

Distinct architectural style of Kashmir

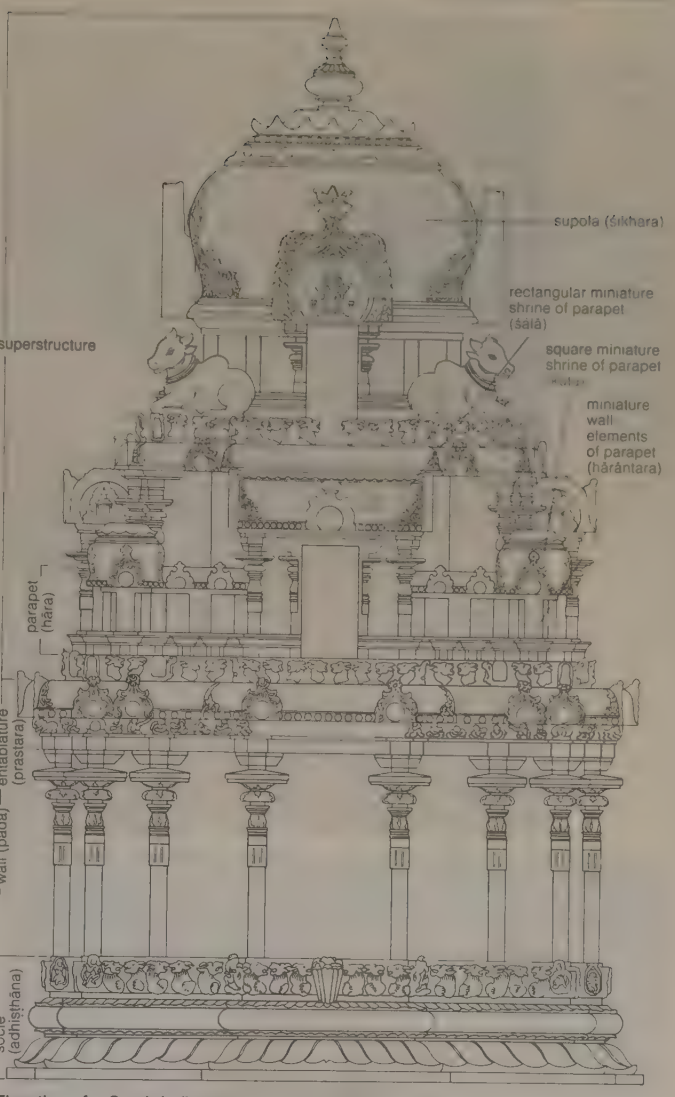
Medieval temple architecture: North Indian style of Kashmir. The architectural style of the Kashmir region is quite distinct: unlike other regions, in which the sanctum usually has a *latina* or *śekharī* spire, the roof of the Kashmir sanctum is of the *phāmsanā* type, with eaves raised in two stages. The greatest example to survive is the ruined Sun Temple at Mārtaṇḍ (mid-8th century), which, though its *śikhara* is missing, gives a good idea of the characteristic features of the style. The temple is placed in a rectangular court enclosed by a series of columns. Access to the court is through an imposing entrance hall, the walls of which have doorways with gabled pediments and a trefoil (shaped like a trifoliate leaf) recess. The Avantisvāmī temple of the mid-9th century, now quite ruined, must have been similar, though much more richly ornamented. The style continued up to the 12th century; the Rilhaṇeśvara temple at Pāndreṇthan is a comparatively well-preserved example of this period.

Medieval temple architecture: South Indian style. The home of the South Indian style, sometimes called the drāviḍa style, appears to be the modern state of Tamil Nadu; examples, however, are found all over southern India, particularly in the adjoining regions of Karnataka and Andhradeśa, now largely covered by the states of Karnataka and Andhra Pradesh. Both Andhradeśa and Karnataka developed variants, particularly Karnataka, which evolved a distinct manner, basically South Indian but with features of North Indian origin. The Karnatic style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora is the most famous example.

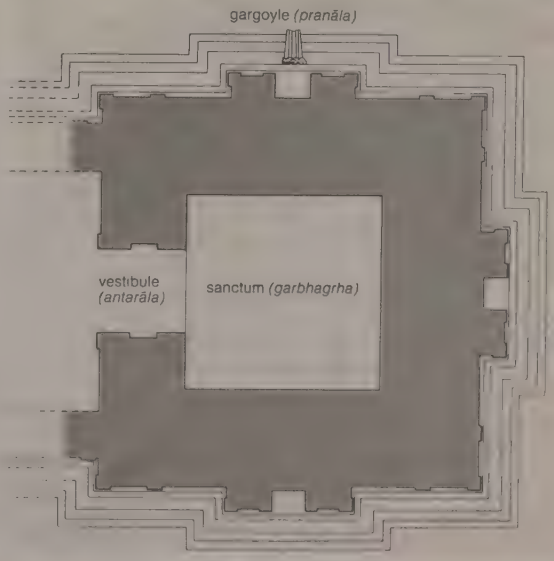
A typical South Indian temple consists of a hall and a square sanctum that has a superstructure of the *kūṭina* type. Pyramidal in form, the *kūṭina* spire consists of stepped stories, each of which simulates the main story and is conceived as having its own "wall" enclosed by a parapet. The parapet itself is composed of miniature shrines strung together: square ones (called *kūṭa*s) at the corners and rectangular ones with barrel-vault roofs (called *śālā*s) in the centre, the space between them connected by miniature wall elements called *hārāntara*s. (Conspicuous in the early temples, these stepped stories of the superstructure with their parapets became more and more ornamental, so that in the course of time they evolved into more or less decorative bands around the pyramidal superstructure.) On top of the stepped structure is a necking that supports a solid dome, or cupola (instead of the North Indian grooved disc), which in turn is crowned by a pot and finial. The walls of the sanctum rise above a series of moldings, constituting the foundation block, or socle (*adhiṣṭhāna*), that differ from North Indian temples; and the surface of the walls does not have the prominent offsets seen in North Indian temples but is instead divided by pilasters. In the Karnatic version, particularly from the late 10th century onward (sometimes called the *vesara* style), this arrangement of the superstructure is loaded with decoration, thus considerably obscuring the component elements. At the same time, these elements—particularly the central offset with its subsidiaries that carry *candraśālā* motifs—are so manipulated that they tend to form distinct vertical bands, in this respect closely recalling the *śikhara*s of northern India.

The design of the hall-temple roofed by a barrel vault, popular in the centuries before and after Christ, was adopted in southern India for the great entrance buildings, or *gopura*s, that give access to the sacred enclosures in which the temples stand. Relatively small and inconspicuous in the early examples, they had, by the mid-12th century, outstripped the main temple in size.

Medieval temple architecture: South Indian style of Tamil Nadu (7th–18th century). The early phase, which, broadly speaking, coincided with the political supremacy of the Pallava dynasty (*c.* 650–893), is best represented by the important monuments at Mahābalipuram. Besides a fine group of small cave temples (early 7th century), among the earliest examples of their type in southern India, there are



Elevation of a South Indian temple with the *kūṭina* type of superstructure.



Plan of the sanctum of a South Indian temple.

The Shore Temple

here several monolithic temples carved out of the rock, the largest of which is the massive three-storied Dharmarājaratha (*c.* 650). The finest temple at this site and of this period is an elegant complex of three shrines called the Shore Temple (*c.* 700), not cut out of rock but built of stone. The Tālapuriśvara temple at Panamalai is another excellent example. The capital city of Kānchipuram also possesses some fine temples—for example, the Kailāsanātha (dating a little later than the Shore Temple), with its stately superstructure and subsidiary shrines attached to the walls. The enclosure wall has a series of small shrines on all sides and a small *gopura*. Another splendid temple at Kānchipuram is the Vaikuṇtha Perumāl (mid-8th century), which has an interesting arrangement of three sanctums, one above the other, encased within the body of the superstructure.

The 9th century marked a fresh movement in the South Indian style, revealed in several small, simple, but most elegant temples set up during the ascendancy of the Cōḷa and other contemporary dynasties. Most important of a large number of unpretentious and beautiful shrines that dot the Tamil countryside are the Vijayālaya Cōḷīśvara temple at Nārttāmalai (mid-9th century), with its circular sanctum, spherical cupola, and massive, plain walls; the twin shrines called Agastyīśvara and Cōḷīśvara, at Kīḻaiyūr (late 9th century); and the splendid group of two temples (originally three) known as the Mūvarkovil, at Koḍumbāḷūr (*c.* 875).

These simple beginnings led rapidly (in about a century) to the mightiest of all temples in the South Indian style, the Bṛhadīśvara, or Rājarājeśvara, temple, built at the Cōḷa capital of Thanjāvūr. A royal dedication of Rājarāja I, the temple was begun around 1003 and completed about seven years later. The main walls are raised in two stories, above which the superstructure rises to a height of 190 feet (60 metres). It has 16 stories, each of which consists of a wall with a parapet of shrines carved in relatively low relief. The great temple at Gaṅgaikoṇḍacōḷapuram, built (1030–40) by the Cōḷa king Rājendra I, is somewhat smaller than the Bṛhadīśvara; but the constituent elements of its superstructure, whose outline is concave, are carved in bolder relief, giving the whole a rather emphatic plasticity. The Airāvateśvara (1146–73) and Kampahareśvara (1178–1223) temples at Dārāsuram and Tribhuvanam follow the tradition of the 11th century but are smaller and considerably more ornate. They bring to a close a great phase of South Indian architecture extending from the 11th to the 13th century.

P Chandra

Mūvarkovil at Koḍumbāḷūr, Tamil Nadu, India, *c.* AD 875.

From the middle of the 12th century onward, the *gopuras*, or entrance buildings, to temple enclosures began to be greatly emphasized. They are extremely large and elaborately decorated with sculpture, quite dominating the architectural ensemble. Their construction is similar to that of the main temple except that they are rectangular in plan and capped by a barrel vault rather than a cupola, and only the base is of stone, the superstructure being made of brick and plaster. Among the finest examples are the Sundara Pāndya *gopura* (13th century) of the Jambukeśvara temple at Tiruchchirāppalli and the *gopuras* of a great Śiva temple at Chidambaram, built largely in the 12th–13th century. Even larger *gopuras*, if not of such fine quality, continued to be built up to the 17th century. Such great emphasis was placed on the construction of *gopuras* that enclosure walls, which were not really necessary, were especially built to justify their erection. In the course of time several walls and *gopuras* were successively built, each enclosing the other so that at the present day one often has to pass through a succession of walls with their *gopuras* before reaching the main shrine. A particularly interesting example is the Ranganātha temple at Srīrangam, which has seven enclosure walls and numerous *gopuras*, halls, and temples constructed in the course of several centuries. The *gopuras* of the Mīnākṣī temple at Madurai are also good representative examples of this period.

In addition to the *gopuras*, temples also continued to be built. Although they never achieved colossal size, they are often of very fine workmanship. The Subrahmaṇya temple of the 17th century, built in the compound of the Bṛhadīśvara temple at Thanjāvūr, indicates the vitality of architectural traditions even at this late date.

Rock-cut cave temples of South India

Medieval temple architecture: South Indian style of Karnataka. The early phase, as in Tamil Nadu, opens with the rock-cut cave temples. Of the elaborate and richly sculptured group at Bādāmi, one cave temple is dated 578, and two cave temples at Aihole are early 8th century. Among structural temples built during the rule of the Cālukyas of Bādāmi are examples in the North Indian style; but, because the Karnataka region was more receptive to southern influences, there are a large number of examples that are basically South Indian with only a few North Indian elements. The Durgā temple (*c.* 7th century) at Aihole is apsidal in plan, echoing early architectural traditions; the northern *latina śikhara* is in all probability a later addition. The Mālegitti Śivālaya temple at Bādāmi (early 8th century), consisting of a sanctum, a hall with a parapet of *śālās* and *kūṭas* (rectangular and square miniature shrines), and an open porch, is similar to examples in Tamil Nadu. The Virūpākṣa at Pattadkal (*c.* 733–746) is the most imposing and elaborate temple in the South Indian manner. It is placed within an enclosure, to which access is through a *gopura;* and the superstructure, consisting of four stories, has a projection in the front, a feature inspired by the prominent projections, or *śukanāsā,* of North Indian temples. Belonging to the 9th century is the triple shrine (the three sanctums sharing the same *maṇḍapa,* or hall) at Kambaḍahalli and the extremely refined and elaborately carved Bhoganandīśvara temple at Nandi. The Chāvuṇḍarāyabasti (*c.* 982–995) at Śravaṇa-Beḷgoḷa is also an impressive building, with an elegant superstructure of three stories.

With the 10th century, the Karnatic idiom begins to show an increasing individuality that culminates in the distinctive style of the 12th century and later. The Kalleśvara temple at Kukkanūr (late 10th century) and a large

Jaina temple at Lakkundi (*c.* 1050–1100) clearly demonstrate the transition. The superstructures, though basically of the South Indian type, have offsets and recesses that tend to emphasize a vertical, upward movement. The Lakkundi temple is also the first to be built of chloritic schist, which is the favoured material of the later period and which lends itself easily to elaborate sculptural ornamentation. With the Mahādevā temple at Ittagi (*c.* 1112) the transition is complete, the extremely rich and profuse decoration characteristic of this shrine being found in all work that follows. Dating from the reign of the Hoysaḷa dynasty (*c.* 1141) is a twin Hoysaḷeśvara temple at Halebīd, the capital city. The sanctums are stellate in form but lack their original superstructures. The pillars of the interior are lathe-turned in a variety of fanciful shapes. The exterior is almost totally covered with sculpture, the walls carrying the usual complement of images; the base, or socle, is decorated with several bands of ornamental motifs and a narrative relief. Among other temples that were constructed in this style, the most important are the Chenna Keśava temple at Belūr (1117), the Amṛteśvara temple at Amritpur (1196), and the Keśava temple at Somnāthpur (1268).

Strong traditions of cave architecture in Mahārāshtra

Medieval temple architecture: South Indian style of Mahārāshtra, Andhradeśa, and Kerala. The traditions of cave architecture are stronger in Mahārāshtra than in any other part of India; there, great shrines were cut out of rock right up to the 9th century AD and even later. Of those belonging to the early phase, the most remarkable is a temple at Elephanta (early 6th century); equally impressive are numerous temples at Ellora (6th–9th centuries). The Karnatic version of the South Indian style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora, erected in the reign of the Rāṣṭrakūṭā Krishna I (8th century), is its most stupendous achievement. The entire temple is carved out of rock and is over 100 feet (30 metres) high. It is placed in a courtyard, the three sides of which are carved with cells filled with images; the front wall has an entrance *gopura*. The tall base, or plinth, is decorated with groups of large elephants and griffins, and the superstructure rises in four stories. Groups of important temples in the southern style are also found in the Andhra country, notably at Biccavolu, ranging in date from the 9th to the 11th centuries. The 13th-century temples at Palampet are the counterparts of the elaborate Karnatic style of the same period, but without its overpowering elaboration. The temples of Kerala represent an adaptation of the South Indian style to the great main fall of this region and are provided with heavy sloping roofs of stone that imitate timber originals required for draining away the water.

Islāmic architecture in India: period of the Delhi and provincial sultanates. Although the province of Sind was captured by the Arabs as early as 712, the earliest examples of Islāmic architecture to survive in the subcontinent date from the closing years of the 12th century; they are located at Delhi, the main seat of Muslim power throughout the centuries. The Qūwat-ul-Islām mosque (completed 1196), consisting of cloisters around a courtyard with the sanctuary to the west, was built from the remains of demolished temples. In 1198 an arched facade (*maqṣūrah*) was built in front to give the building an Islāmic aspect, but its rich floral decoration and corbelled (supported by brackets projecting from the wall) arches are Indian in character. The Quṭb Mīnār, a tall (288 feet high), fluted tower provided with balconies, stood outside this mosque. The Aṛhāi-dīn-

Quṭb Mīnār and the Qūwat-ul-Islām mosque at Delhi, *c.* AD 1196.
P Chandra

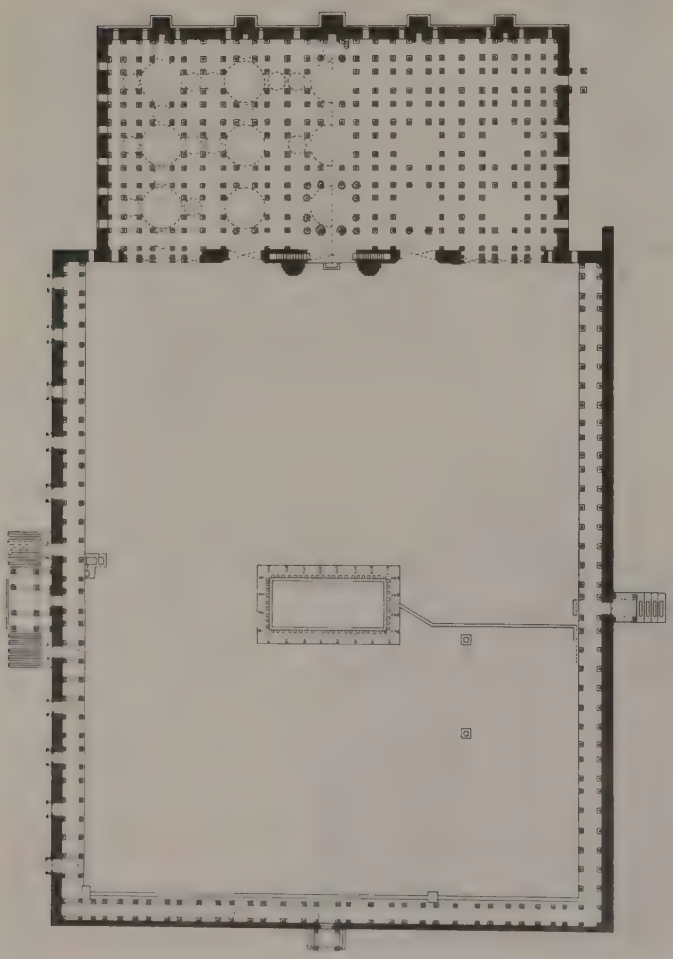
P Chandra

Tomb and palace of Fīrūz Shāh (Hawz-e Khāss) at Delhi, *c.* 1380.

kā-jhompṛā mosque (*c.* 1119), built at Ajmer, was similar to the Delhi mosque, the *maqṣūrah* consisting of engrailed (sides ornamented with several arcs) corbel arches decorated with greater restraint than the Quṭb example. The earliest Islāmic tomb to survive is the Sultān Gharī, built in 1231, but the finest is the tomb of Iltutmish, who ruled from 1211 to 1236. The interior, covered with Arabic inscriptions, in its richness displays a strong Indian quality. The first use of the true arch in India is found in the ruined tomb of Balban (died 1287). From 1296 to 1316 ʿAlāʾ-ud-Dīn Khaljī attempted to expand the Qūwat-ul-Islām mosque, which already had been enlarged in 1230, to three times its size; but he was unable to complete the work. All that has survived of it is the Alai Darwāzah, a beautiful entrance.

14th-century style impoverished and austere

In contrast to this early phase, the style of the 14th century at Delhi, ushered in by the Tughluq dynasty, is impoverished and austere. The buildings, with a few exceptions, are made of coarse rubble masonry and overlaid with plaster. The tomb of Ghiyās-ud-Dīn Tughluq (*c.* 1320–25), placed in a little fortress, has sloping walls faced with panels of stone and marble. Also to be ascribed to his reign is the magnificent tomb of Shāh Rukn-e ʿĀlam at Multān in Pakistan, which is built of brick and faced with exquisite tile work. The Koṭla Fīrūz Shāh (1354–70), with its mosques, palaces, and tombs, is now in ruins but represents the major building activity of Fīrūz Shāh, who took a great interest in architecture. Many mosques and tombs of this period and of the 15th century are found in Delhi and its environs; the most notable of them are the Begampur and Khiṛkī mosques and an octagonal tomb of Khān-e Jahān Tilangānī. In the early 16th century, Shēr Shāh Sūr refined upon this style, the Qalʿah-e Kuhnah Masjid and his tomb at Sasarām (*c.* 1540) being the finest of a series of distinguished works that were created during his reign.



Plan of the Jāmiʿ Masjid, Ahmadābād, Gujarāt.

The provinces, which gradually became independent sultanates, did not lag behind in architectural activity. In West Bengal, at Pandua, is the immense Ādīna Masjid (1364–69), which utilized remains of Indian temples. In Jaunpur, Uttar Pradesh, are a group of elegant mosques, notably the Aṭalā Masjid (1377–1408) and the Jāmiʿ Masjid (*c.* 1458–79), characterized by *maqṣūrah*s that have the aspect of imposing gateways. The sultans of Mālwa built elegant structures at Māndu and at Chanderi in the middle of the 15th century. The sultanate of Gujarāt is notable for its great contribution to Islāmic architecture in India. The style, which is basically indigenous, reinterprets foreign influences with great resourcefulness and confidence, producing works notable for their integrity and unity. The city of Ahmadābād (Ahmedabad) is full of elegant buildings; the Jāmiʿ Masjid (*c.* 1424), for example, is a masterly exposition of the style. Fine examples dating from the second half of the 15th century are the small but exquisite mosques of Muḥāfiz Khān (1492) and Rānī Sabraʾi (1514) at Ahmadābād and the handsome Jāmiʿ Masjid at the city of Chāmpāner.

The Deccan was another great centre, but in contrast to Gujarāt it took little from the indigenous building traditions. Among the earliest works is the Jāmiʿ Masjid at Gulbarga (1367), with its extraordinary cloisters consisting of wide arches on low piers, producing a most solemn effect. The city of Bīdar possesses many remains, including a remarkable series of 12 tombs, the most elaborate of which is that of ʿAlā-ud-Dīn Aḥmad Bahmanī (died 1457), which has extremely fine decorations in coloured tile. Some of the finest examples of Islāmic architecture in the Deccan, however, are in Bijāpur. The most important buildings of this city are the great Jāmiʿ Masjid (begun in 1558) with its superb arched cloisters; the ornate Ibrāhīm Rawẓa; and the Gōl Gunbad (built by Muḥammad ʿĀdil Shāh), a tomb of exceptional size and grandeur, with one of the largest domes in existence.

The Hindu kingdoms that managed to retain varying degrees of independence during the period of Islāmic supremacy also produced important works. These structures naturally bore the imprint of what survived of traditional Indian architecture to a greater extent than did those monuments patronized by Muslims. Among the Hindu structures of this period are the extensive series of palaces, all in ruin, built by Rāṇā Kumbhā (*c.* 1430–69) at Chitor, and the superb Mān Mandir palace at Gwalior (1486–1516), a rich and magnificent work that exerted considerable influence on the development of Mughal architecture at Fatehpur Sīkrī.

Striking revival of Islāmic architecture

Islāmic architecture in India: Mughal style. The advent of the Mughal dynasty marks a striking revival of Islāmic architecture in northern India: Persian, Indian, and the various provincial styles were successfully fused to produce works of unusual refinement and quality. The tomb of Humāyūn, begun in 1564, inaugurates the new style. Built entirely of red sandstone and marble, it shows considerable Persian influence. The great fort at Āgra (1565–74) and the city of Fatehpur Sīkri (1569–74) represent the building activities of the emperor Akbar. The former has the massive so-called Delhi gate (1566) and lengthy and immense walls carefully designed and faced with dressed stone throughout. The most important achievements, however, are to be found at Fatehpur Sīkri; the Jāmiʿ Masjid (1571), with the colossal gateway known as the Buland Darwāza, for example, is one of the finest mosques of the Mughal period. Other notable buildings include the palace of Jodhā Bāī, which has a strongly indigenous aspect; the exquisitely carved Turkish Sultānā's house; the Pānch-Maḥal; the Dīvān-e ʿĀmm; and the so-called hall of private audience. Most of the buildings are of post and lintel construction, arches being used very sparingly. The tomb of the emperor, at Sikandarā, near Āgra, is of unique design, in the shape of a truncated square pyramid 340 feet (103 metres) on each side. It consists of five terraces, four of red sandstone and the uppermost of white marble. Begun about 1602, it was completed in 1613, during the reign of Akbar's son Jahāngīr. Architectural undertakings in this emperor's reign were not very ambitious, but there are fine buildings,

Tomb of 'Isā Khān at Delhi, 1547.
P Chandra

chiefly at Lahore. The tomb of his father-in-law I'timād-ud-Dawla, at Āgra, is small but of exquisite workmanship, built entirely of delicately inlaid marble. The reign of Shāh Jahān (1628–58) is as remarkable for its architectural achievements as was that of Akbar. He built the great Red Fort at Delhi (1639–48), with its dazzling hall of public audience, the flat roof of which rests on rows of columns and pointed, or cusped, arches, and the Jāmi' Masjid (1650–56), which is among the finest mosques in India. But it is the Tāj Mahal (*c.* 1632–*c.* 1649), built as a tomb for Queen Mumtāz Maḥal, that is the greatest masterpiece of his reign. All the resources of the empire were put into its construction. In addition to the mausoleum proper, the complex included a wide variety of accessory buildings of great beauty. The marble mausoleum rises up from a tall terrace (at the four corners of which are elegant towers, or minārs) and is crowned by a graceful dome.

P Chandra

Buildings at Fatehpur Sīkri, Uttar Pradesh, India, *c.* 1571.

Other notable buildings of the reign of Shāh Jahān include the Motī Masjid (*c.* 1648–55) and the Jāmi' Masjid at Āgra (1548–55).

Architectural monuments of the reign of Aurangzeb

Architectural monuments of the reign of Aurangzeb represent a distinct decline; the tomb of Rābī'ah Begam at Aurangābād, for example (1679), is a poor copy of the Tāj Mahal. The royal mosque at Lahore (1673–74) is of much better quality, retaining the grandeur and dignity of earlier work; and the Motī Masjid at Delhi (1659–60) possesses much of the early refinement and delicacy. The tomb of Ṣafdar Jang at Delhi (*c.* 1754) was among the last important works to be produced under the Mughal dynasty and had already lost the coherence and balance characteristic of mature Mughal architecture.

European traditions and the modern period. Buildings imitating contemporary styles of European architecture, often mixed with a strong provincial flavour, were known in India from at least the 16th century. Some of this work was of considerable merit, particularly the baroque architecture of the Portuguese colony of Goa, where splendid buildings were erected in the second half of the 16th century. Among the most famous of these structures to survive is the church of Bom Jesus, which was begun in 1594 and completed in 1605.

The 18th and 19th centuries witnessed the erection of several buildings deeply indebted to Neoclassic styles; these buildings were imitated by Indian patrons, particularly in areas under European rule or influence. Subsequently, attempts were made by the British, with varying degrees of success, to engraft the neo-Gothic and also the neo-Saracenic styles onto Indian architectural tradition. At the same time, buildings in the great Indian metropolises came under increasing European influence; the resulting hybrid styles gradually found their way into cities in the interior. In recent years an attempt has been made to grapple with the problems of climate and function, particularly in connection with urban development. The influence of the Swiss architect Le Corbusier, who worked on the great Chandigarh project, involving the construction of a new capital for Punjab, in the early 1950s, and that of other American and European masters has brought about a modern architectural movement of great vitality, which is in the process of adapting itself to local requirements and traditions.

Sculpture. On the Indian subcontinent, sculpture seems to have been the favoured medium of artistic expression. Even architecture and the little painting that has survived from the early periods partake of the nature of sculpture. Particularly is this true of rock-cut architecture, which is

Church of Bom Jesus at Velha Goa, India, 1594–1605.
P. Chandra

often little more than sculpture on a colossal scale. Structural buildings are also profusely adorned with sculpture that is often inseparable from it. The close relationship between architecture and sculpture has to be taken into account when considering individual works that, even if complete in themselves, are also fragments belonging to a larger context. Indian sculpture, particularly from the 10th century onward, thus cannot be studied in isolation but must be considered as part of a larger entity to the total effect of which it contributes and from which it in turn gains meaning.

The subject matter of Indian sculpture is almost invariably religious. This does not mean that it cannot be understood as a work of art apart from its religious significance; but, at the same time, an understanding of its motivation and intent enriches one's appreciation. Much of what is represented is the recounting of legend and myth, particularly in the two centuries before Christ, when narrative relief was much in vogue. The work at this time, didactic and edificatory in intent, generally expresses itself in forms that are surprisingly earthy and sensuous. The anthropomorphic representation of the Buddha is avoided, and the subsidiary gods and goddesses are very much creatures of this earth. The Buddha image formulated around the 1st century AD is not what one would expect of the meditative, compassionate, Master of the Law; he is presented rather as an energetic, earthy being radiating strength and power.

Buddhist influences

The foundations of traditional Hindu imagery were also laid about the same time that the Buddha image was first formulated: images with several arms, and sometimes heads, representing the Indian mind's attempt to define visually the infiniteness of divinity. In subsequent periods the image with many arms became a commonplace in Hindu, Buddhist, and Jaina iconography. Although the various pantheons expanded, they continued to share features of common derivation, expressing the belief that beyond the phenomenal multiplicity of forms lay the unity of the Godhead.

In addition to the major religions, there has always existed in India a substratum of folk beliefs and cults dedicated to the worship of powers that preside over the operation of the life processes of nature. These fertility cults, best expressed in the worship of the male and female divinities *yakṣas* and *yakṣīs*, played an important part in the development of Indian art. Among the perennial motifs that spring from the cults, those expressing life and

Paolo Koch—Rapho/Photo Researchers

The Legislative Assembly chambers of the states of Haryana and Punjab in Chandigarh, India, designed by Le Corbusier, 1952.

abundance—such as the lotus, the pot overflowing with vegetation, water, or the like, the tree, the amorous couple, and above all the *yakṣas* and *yakṣīs* themselves—are most significant. The images of these divinities, in particular, are the source of a great deal of artistic imagery and played a leading part in the development of iconographic types such as the images of the Buddha, the goddess Śrī, and other divinities. The maternal as the ideal of female beauty, which is manifested artistically in the emphasis on full breasts and wide hips, can be traced to the same beliefs. The very richness and exuberance of much Indian art is an expression of the view of life that equates beauty with abundance.

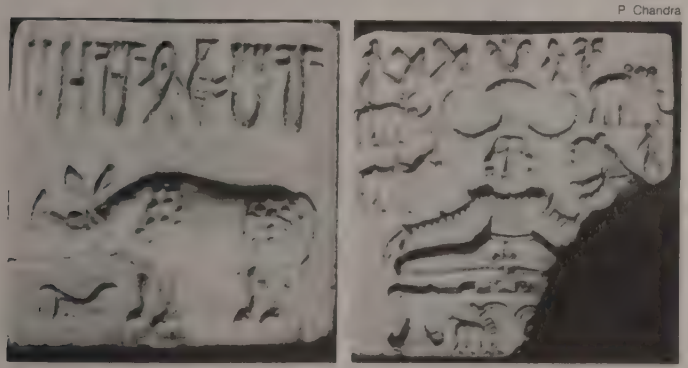
It is difficult to generalize about the style of a sculptural tradition that extended over a period of almost 5,000 years, but it is nevertheless clear that the distinguishing quality of Indian sculpture is its emphatic plasticity so obvious in Sānchi I and Mathurā sculpture from the 1st–3rd century AD. Forms are seen as swelling from within in response to the power of an inner life, the sculptor's function being to make these more manifest. At the same time a vision of form that is carved from without rather than modelled from within is also present, as for example at Bhārhut. The history of much of Indian sculpture, marked by periods of high achievement bursting with creativity followed by periods in which the potentialities so postulated are gradually worked out, is essentially the interaction of these two dominant tendencies.

Indus Valley civilization (c. 2500–1800 BC). Sculpture found in excavated cities consists of small pieces, generally terra-cotta objects, soapstone, or steatite, seals carved for the most part with animals, and a few statuettes of stone and bronze. The terra-cotta figurines are summarily modelled and provided with elaborate jewelry, which was fashioned separately and applied to the surface of the piece. Most of the work is simple, but a small group of human heads with horns are very sensitively modelled. Animal figures are common, particularly bulls, which are often carved with a sure understanding of their bulky, massive form. This plastic quality is also found in the humped bulls engraved on steatite seals, where the modelling is more refined and sensitive. A humpless beast, generally called a "unicorn," is another favourite animal, but it is frequently quite stylized. In addition to bisons, elephants, rhinoceroses, and tigers, seals are carved with images of apparent religious significance, often strongly pictographic.

Terra-cotta sculpture and seals

The terra-cotta sculpture and the seals both show two clear and distinct stylistic trends, one plastic and sensuous, the other linear and abstract. These appear during the same period and are also seen in the small group of stone and bronze sculptures that date from this period (National Museum, New Delhi). Of extraordinarily full and refined modelling is a fragmentary torso from Harappā, barely four inches (10 centimetres) high but of imposing monumentality; the same feeling for massive form is present in a lesser known bronze buffalo. A jaunty bronze dancing girl with head tilted upward (about 4½ inches [11 centimetres] high), from Mohenjo-daro, and a headless figure of a male dancer from Harappā, shoulders twisted in a circular movement, clearly demonstrate, in the attenuated and wiry tension of their forms, the second component of Indus Valley art. Of great interest is a famous bearded figure from Mohenjo-daro wearing a robe decorated with a pattern composed of trefoil motifs. The tight, compressed shape of the body and the expansive modelling of the head demonstrate that the two aspects of form revealed in Indus Valley art were not compartmentalized but interacted with each other. This can also be seen in the interplay of modelled form and textured surface frequently found in works produced by this civilization.

P. Chandra



Steatite seals of the Indus Valley civilization (c. 2300–c. 1750 BC). In the National Museum of India, New Delhi.

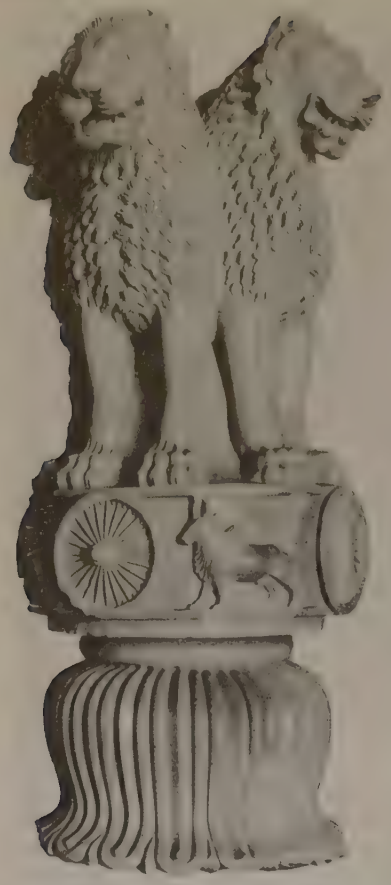
Maurya period (c. 3rd century BC). Little is known of Indian art in the period between the Indus Valley civilization and the reign of the Maurya emperor Aśoka. When sculpture again began to be found, it was remarkable for its maturity, seemingly fully formed at birth. The most famous examples are great circular stone pillars, products of Aśoka's imperial workshop, found over an area stretching from the neighbourhood of Delhi to Bihār. Made of fine-grained sandstone quarried at Chunār near Vārānasi (Benares), the monolithic shafts taper gently toward the top. They are without a base and, in the better preserved examples, are capped by campaniform lotus capitals supporting an animal emblem. The entire pillar was carefully burnished to a bright lustre commonly called the "Maurya polish." The most famous of these monuments is the lion capital at Sārnāth, consisting of the front half of four identical animals joined back to back. There is a naturalistic emphasis on build and musculature, and the modelling is hard, vigorous, and energetic, stressing physical strength and power. Very similar, if not at the same level of achievement, is the quadruple lion capital at Sānchi. Single lions are found at Vaiśālī (Bakhra), Rāmpurvā, and Lauriya Nandangarh. The Vaiśālī pillar is heavy and squat, and the animal lacks the verve of the other animals—features, according to some, designating it as an early work, executed before the Maurya style attained its maturity. By contrast, the Rāmpurvā lion, finished with painstaking and concise artistry, represents the style at its best. His smooth, muscled contours, wiry sinews, rippling, flamboyant mane, and alert stance reveal the work of a superior artist. An example at Lauriya Nandangarh is interesting because the pillar and the lion are both complete and in their original place, giving a clear idea of the column as it appeared to its contemporaries.

The Rāmpurvā lion

The lion was the animal most often represented, but figures of elephants and bulls are also known. At Dhauli in Orissa, the fore part of an elephant is carved out of rock on a terrace above a boulder that carries several of Aśoka's edicts. The modelling here is soft and gentle, and the plump, fleshy qualities of the young animal's body, seen as emerging from the rock, are suffused with warmth and natural vitality. Since the contrast with the rather formal, heraldic lions could not be more complete, the sculpture clearly testifies to the simultaneous existence of a style different from that of the lion capitals. The style might very well represent the indigenous tradition of plastic form that appears consistently in later art and also in some of the animal capitals made in the imperial atelier, notably the damaged elephant that once crowned the pillar at Saṅkisa and, above all, the splendid bull from Rāmpurvā. In this great work of art, the two opposing concepts of form merge in a work of harmonious power. The pronounced naturalism comes from the same source as do the lions, but the tense line and hard modelling yield to a form that wells from within and at the same time is given stability and strength by a vision imposed from without.

The sudden appearance of Maurya art with seemingly no tradition behind it has led to speculation that it was the creation of foreign artists, either Achaemenian or Hellenistic. Persian influence, particularly in the lotus capitals and the figures of lions can hardly be denied, but what is remarkable is the drastic reinterpretation of alien forms by Indian artists. This is a process that is repeatedly seen in the history of Indian art.

Besides the animal sculpture, some human figures, more or less life size, can also be assigned to the Maurya period, though scholarly opinion is by no means unanimous on the point. Among the most important are three images discovered at Patna (ancient Pāṭaliputra, the Maurya



Lion capital from Sārnāth, Uttar Pradesh, India, Chunār sandstone, mid-3rd century BC. Height 2.13 m.
P. Chandra

capital), two of which are representations of *yakṣa*s, the popular male divinities associated with cults of fertility, and the third, found at Dīdarganj (a section of Patna), a representation of a *yakṣī*, or female divinity. Stylistically the images are very similar. The standing *yakṣa*s (Indian Museum, Calcutta) are powerful creatures; the ponderous weight of their bodies, together with a certain refined appreciation of the soft flesh, is admirably rendered. The Dīdarganj *yakṣī* (Patna Museum), a masterpiece, displays the Indian ideal of female beauty, the heavy hips and full breasts strongly emphasizing the maternal aspect. In a nude torso discovered at Lopanipur, the sophisticated and sensitive treatment of the surfaces and the gentle blending planes that avoid all harsh accents produce a work of much refinement.

Small stone discs

Small stone discs (also called ring stones because several of them are perforated in the centre), found from Taxila to Patna, are clearly connected with the cult of a nude mother goddess. They represent Maurya sculpture on a smaller and more intimate scale but characterized by the same refined and exquisite workmanship. They are executed in bas-relief, which became the favourite form of sculpture in the subsequent period.

The terra-cotta art of the Maurya period is best represented by a substantial group of figurines, modelled for the most part, the clay sculptor performing work in his medium at the same level as the artist working in stone. Patna has yielded a large number of such works, but examples are found throughout the Gangetic Plain. The clothing and jewelry on the figurines are heavy and elaborate, the modelling, particularly of the head, is sensitive, and the expression is often one of great charm and refinement. There are also more archaic examples, distinguished by flat bodies, enormous hips, and modelled heads and breasts.

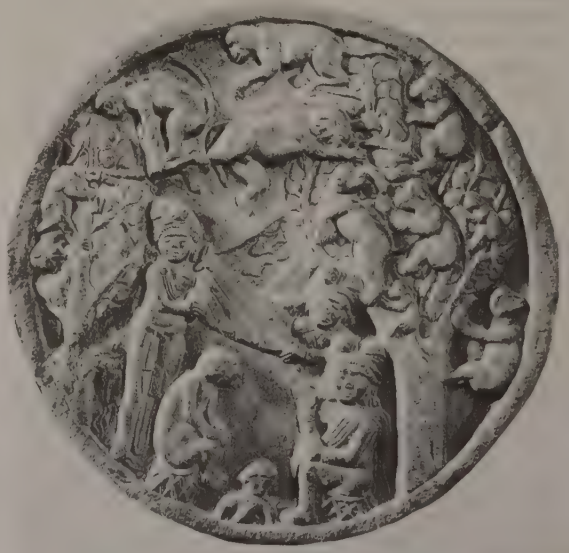
Indian sculpture in the 2nd and 1st centuries BC. The Maurya Empire collapsed in the early years of the 2nd century BC, and with it passed the art with which it was intimately related. The sculpture that is found throughout India from the middle of the 2nd century BC is startlingly different, but the process by which this change took place in a relatively short period of time is not fully understood. Several schools, sharing common features but nevertheless possessing distinct individual characteristics, are known to have existed. The history of the schools of northern India is somewhat obscure, largely due to the great destruction wrought in the Gangetic heartland; but there appears to have flourished there and in adjacent areas a school of great importance represented by the remains discovered at Bhārhut, Sānchi, Mathurā, and Buddh Gaya. Western India had its own school, as revealed in the sculptures decorating the cave temples, notably those of Bhājā, Pītalkhorā, and Kārli. In the southeast, the important school of Andhradeśa flourished in the Krishna River Valley at Amāravatī, Jaggayyapeta, and associated sites; and in eastern India, what is now the modern state of Orissa, made its contribution in the rock-cut sculptures at Udayagiri-Khandagiri. The distinctive schools, though spread over a subcontinent, were not isolated from each other. The contacts fostered by a flourishing trade and by the constant movement of pilgrims were always very close, and it was never long before developments in one part of India were echoed in another.

Judging from extant remains, artists of the earlier period (*c.* 3rd century BC) preferred figures carved in the round, relief sculpture being quantitatively quite insignificant. By contrast, it was sculpture in low relief that was favoured in the first two centuries before Christ; the earlier tradition was not quite forgotten, but figures carved in the round are relatively few. Although there is no stylistic difference, relief sculpture is here considered first according to the various regional schools, and sculpture in the round is treated separately.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of northern and central India. Among the most important, and perhaps the earliest, remains in northern India are reliefs from the great *stūpa* at Bhārhut, dating approximately to the middle of the 2nd century BC. The work, suggesting a style imitating wooden sculpture, is characterized by essentially cubical forms, flat planes that meet at sharp angles, and very elaborate and precisely detailed ornamentation of surfaces. Most of the sculpture was confined to the railing of the *stūpa*. Some of the supporting posts bear large image of *yakṣa*s and *yakṣī*s of popular religion, now clearly pressed into the service of Buddhism, while most of the others are decorated with medallions in the centre and crescent-shaped motifs, or lunates, at the top and bottom, all filled with

Reliefs from the *stūpa* at Bhārhut

P. Chandra



The monkey chief and the king of Vārānasi, red sandstone relief from Bhārhut, Madhya Pradesh, India, mid-2nd century BC. In the Indian Museum, Calcutta.

lotus motifs. Some medallions contain amorous couples, the overflowing pot, the goddess Śrī standing on lotuses while being ceremonially bathed by elephants and other symbols of abundance; still others contain the earliest illustrations of events in the Buddha's life and of narratives of his former incarnations as related in the *Jātaka* tales (a collection of tales about the Buddha). Although compositions are crowded, great economy of expression is evident because the artist confines himself to the representation of essentials. Figures are often carved in horizontal rows, sometimes asymmetrically, adapting themselves awkwardly to the circular space of the medallion. Continuous narrative, in which events succeeding in time are shown in the same space, is often resorted to—the first occurrence of what was to become a favourite narrative technique. There is no attempt at establishing any interrelationship, psychological or compositional, between the various figures, each of which is strictly confined within its own space. The faces are masklike, without trace of emotion, lending a solemn and hieratic quality to their expression. Trapped between the background and a frontal plane beyond which they are not allowed to project, the figures are in a sense strictly two-dimensional, more so than in any other style of Indian sculpture. Often, however—particularly in the treatment of animals—the artist is more relaxed, giving glimpses of intimate observation and a natural rendering that anticipates the direction of future development. Like the posts, the top part, or coping, of the stone rail is also carved on both faces; on one of them is a continuous creeper bearing lotus flowers, leaves, and buds; on the other, again the winding stem of a creeper, but bearing other good things of life—such as clothes, jewelry, and fruits—and also scenes illustrating *Jātaka* stories.

Bhārhut is an extremely important monument inasmuch as it seems to mark a new beginning after the refined and naturalistic art of the Maurya Empire. The sophistication, in spite of the archaic, hieratic manner, would indicate that a considerable body of sculptural tradition, particularly in wood, preceded it; but of this no traces have survived. Be that as it may, Bhārhut states for the first time, and at some length, themes and motifs that would henceforth remain a part of Indian sculpture.

Stray finds of sculpture at Mathurā and other sites in modern Uttar Pradesh indicate that the Bhārhut style was spread over a large part of northern India, particularly the region roughly between that city and Vārānasi and Buddh Gaya in the east. A closely related style is also found at Sānchi in eastern Mālava, where a representative example is the sculpture of the railing of Stūpa II. Although the themes and motifs found at Bhārhut occur here, narrative representations are all but absent. The style is almost identical; the stiff and rigid contours are a little softer, but both the scale and richness of Bhārhut are missing.

Principal glory of Sānchi

It is the sculpture of the four gateways (*toraṇas*) of the Great Stūpa (Stūpa I) at Sānchi, however, that is the principal glory of that site, carrying the promise of the Bhārhut style to its fulfillment. The *toraṇas*, four in number, were attached to the plain railing around the middle of the 1st century BC. They consist of square posts with capitals supporting a triple architrave, or molded band, with voluted (turned in the shape of a spiral, scroll-shaped ornament) ends and a top crowned with Buddhist symbols. Bracket figures, in the form of *yakṣīs*, serve as additional supports. All parts of these gates, strongly reminiscent of wooden construction, are covered from top to bottom with the most exquisite sculpture. Subjects and motifs found at Bhārhut are also found here, the same profusely flowering lotus stem and associated motifs, the same compositions with figures basically arranged in horizontal rows, the same love for clear detail; but to all of these are added a truly voluminous sense of form, a smoother and more energetic movement, and a keen appreciation for the forms of nature, all of which endow the sculpture with a naïve and sensuous beauty unparalleled in Indian art.

Departures from the Bhārhut style are particularly striking in the narrative reliefs. Their greater depth, taken together with their crowded composition, results in the background, visible at Bhārhut, being submerged in shadow. The figures, in all their richness and abundance, flow out from the dark ground, secured in place by the frame of the panels. The Bhārhut angular silhouette and the rigid, severe outline of the body yields at Sānchi to a gently swelling plasticity, animated by a soft, breathing quality that molds the contours without strain or tension. There is a pronounced concern with the organization of composition, and the narration is often leisurely and discursive; the artist does not just tell the basic story but also lingers over the details, amplifying them to give a vivid picture of everyday life. The emotional monotone of Bhārhut survives in some Sānchi sculptures, but in others it is superseded by joyous faces and the emotional impact of vivid gesture and movement. Dejection is written large on the faces of the soldiers of Māra's army, who had tried to disturb the Buddha's meditation, as they stagger away from the scene of defeat, and the sensuousness of the amorous scenes is successfully evoked by the tender and intimate gestures of the couples. No longer transfixed in their own space, they turn to look at each other lovingly, responding to each other with a deeply felt understanding.

Long and elaborate bas-reliefs carved on the architraves of the *toraṇas* are the summit of the Sānchi sculptor's art. Among the finest are representations of the wars for the relics, the defeat of Māra, the *Viśvāntara Jātaka*, and the *Ṣaḍḍanta Jātaka*. The compositions are rich and crowded with figures, and are arranged with great skill. Particularly striking is the masterly handling of animals, notably the elephant, whose fleshy body and graceful movement are captured unerringly. Deer, water buffaloes, bulls, monkeys—all of the beasts and birds of the forests—are rendered with a sense of intimacy indicating the artist's sense of the fellowship of man and animal in the world of nature. The lush Indian landscape is often carved with ornamental trees, waterfalls, pools, mountains, and rivers. The Sānchi sculptor also shows a marked preference for architectural settings, filling his compositions with numerous buildings that often provide the spatial context for the action. Entire cities, with surrounding walls, elaborate gate houses, and palatial mansions, are depicted. Depth is achieved by rendering side views, and multiple perspective continues to be the rule.

Images of *yakṣīs*

The several large images of *yakṣīs* serving as brackets supporting the lowermost architraves of the *toraṇas* are unique achievements. Like the same goddesses at Bhārhut, they are shown in association with a tree to which they cling, but the style is remarkably different. The modelling shows a concern for the charms of the body, stressing the tactile nature of its flesh. The heavy jewelry and clothing that conceal the body are drastically reduced, revealing its nudity. The soft, melting sensuousness of the female form is so greatly emphasized that the belly and the folds of flesh at the waist are almost flabby, redeemed only by the smooth, firm breasts and the tender arms and limbs.

By comparison, reliefs adorning the railing around the Mahābodhi temple at Buddh Gaya (of about the same date or a little earlier) are in a somewhat impoverished idiom, lacking the rich proliferation both of Bhārhut and Sānchi. The posts have the usual medallions, lunates filled with lotuses, and reliefs depicting the familiar scenes of Buddhist myth and legend. The artistry of Buddh Gaya, however, is of a lower level of achievement than that at either Bhārhut or Sānchi: the relief is deeper than that at Bhārhut but shallower than that at the Great Stūpa of Sānchi; and crowded compositions are lacking, as are the clear and precise ornament and the rich floral motifs. The Buddh Gaya sculptor, however, though abbreviating even further the iconography of Bhārhut, breaks up, as does the Sānchi sculptor, the spatial isolation that so uncompromisingly separated each individual figure at that site.

The great school of Mathurā, also, seems to have come into existence about the 2nd century BC, though its period of greatest activity falls in the first two centuries after Christ. The city was repeatedly sacked in the course of the centuries, which may account for the paucity of materials, but enough has been discovered to reveal that the style, in its early stages, was very similar to that of Bhārhut, characterized by flat two-dimensional sculpture decorated with abundant and precise ornament. Several fragments

discovered at the site show the gradual stages by which this style evolved, leading to the sculpture of the Great Stūpa at Sānchi on the one hand and to Buddh Gaya on the other.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Andhradeśa. Besides the schools of northern India, a very accomplished style also existed in southeast India; the most important sites are Jaggayyapeta and Amarāvatī, activity at the latter site extending well into the 2nd century AD. The early remains are strikingly similar to those at Bhārhut, the relief generally even shallower and the modelling comparatively flat. In contrast to those found in northern India, the proportions of the human body are elongated; but in its flat, cubical modelling, angular, halting contours, and precise, detailed ornamentation, the style is essentially similar to contemporary work elsewhere, right down to the same conventional clothing and jewelry. The nervous, fluid treatment of surfaces, so characteristic of subsequent Andhra sculpture, is already present here. The preferred material is marble rather than the sandstone invariably used in the north.

The style of the Andhradeśa school developed in a manner consistent with other regions of India, becoming more voluminous and shedding the early rigidity fairly rapidly. A group of sculptures at Amarāvatī are characterized by the same qualities that distinguish the work at the Great Stūpa of Sānchi: full and lissome forms, modelling that emphasizes mass and weight, and sensuously rendered surfaces.

Rock-cut cave temples of the Western Ghāts

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of western India. The numerous rock-cut cave temples in the Western Ghāts are, comparatively speaking, much less profusely adorned with sculpture than remains from other parts of India. The earliest works are undoubtedly the bas-reliefs on a side wall of the porch of a small monastery at Bhājā. They are commonly interpreted as depicting the god Indra on his elephant and the sun god Sūrya on his chariot but are more probably illustrations of the adventures of the mythical universal emperor Māndhātā. What is immediately evident is that these sculptures are not imitations of wooden prototypes, like those at Bhārhut, but, rather, reflect a tradition of terra-cotta sculpture, abundant examples of which are found in northern India and Bengal, where this medium was very popular because of the easy availability of fine clay. The terra-cotta tradition is reflected in the amorphous, spreading forms of Bhājā and in the fine striations used in depicting ornaments and pleated cloth, techniques natural and appropriate to the fashioning of wet clay. The fact that there are some similarities to the Bhārhut style—the stilted postures of the figures and the flat contours of the body, for example—indicates that the beginnings of the western Indian school would also have to be placed about the middle of the 2nd century BC.

The next major group of sculptures in western India have been found at Pītalkhorā. The colossal plinth of a monastery decorated with a row of elephants, the large figures of the door guardians, and several fragments recovered during the course of excavations are among the more important remains. A great proportion of the work represents an advance over the style of Bhājā, though features derived from terra-cotta sculpture continue to be found: the figures are carved in greater depth and volume, but the texture of the drapery, the soft contours of the body, and the high relief of the jewelry, which sometimes gives the impression of having been fashioned separately and then applied, testify to the continuing strength of the terra-cotta tradition. Although the hard line and sharp cutting of some sculpture is reminiscent of the earlier, wood-carving tradition as seen at Bhārhut, the forms are more appropriate to the stone medium. Moreover, the expression is more explicit; and for the first time, both gently smiling and boldly laughing figures of *yakṣa*s appear, as well as the figure of a lover blissfully drunk on wine offered to him by his beloved. These features are also found in the later sculpture of the Great Stūpa at Sānchi and, to a more pronounced extent, in the sculpture of the Mathurā school of the 1st centuries AD—for example, in the happily smiling *yakṣī*s from Bhutesar.

The cave temple at Kondane has, above the entrance hall, four beautiful panels depicting pairs of dancers. The forms retain the robust and full modelling of the more developed sculpture at Pītalkhorā, but to this is added an ease of movement and considerable rhythmic grace. Traces of the terra-cotta tradition are now totally absent; nor do they occur in the next phase, best represented by a group of sculptures found in the rock-cut temples and monasteries at Bedsā and Nāsik and in the *caitya*, or temple proper, at Kārli. Sculpture at all these sites shows

P Chandra

Amorous couple, detail of the *caitya* at Kārli, Mahārāshtra, India.

many affinities to the Great Stūpa at Sānchi and should be approximately contemporary or a little earlier. Easily the most outstanding achievements of this region and period, and for that matter one of the greatest achievements of the Indian sculptor, are the large panels, depicting amorous couples, located in the entrance porch of the Kārli *caitya*. Here the promise of early work achieves its fulfillment, the full weighty forms imbued with a warm, joyous life and a free, assured movement. The resemblance to work at the Great Stūpa of Sānchi is obvious, though these figures at Kārli are on a much larger scale and possess a massiveness and monumentality that is a characteristic of the distinct western Indian idiom.

Sculpture decorating rock-cut monasteries

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Orissa. Sculpture decorating the monasteries cut into the twin hills of Udayagiri and Khandagiri in Orissa represents yet another early Indian local idiom. The work is not of one period but extends over the first two centuries before Christ; the stages of development roughly parallel the styles observed at Sānchi Stūpa No. II, Buddh Gaya, and the Great Stūpa at Sānchi, but they possess, like other regional schools, fairly distinct and individual features. The earliest sculptures are the few simple reliefs found in the Alakāpurī cave, humble works that recall the bas-reliefs of Sānchi Stūpa II. The Mañcapurī, Tatowā Gumphā, and Anantā cave sculptures—particularly the image of Sūrya riding a chariot—are more advanced and resemble work at Buddh Gaya. The forms are heavy and solid and lack the accomplished movement of the later cave sculpture adorning the Rānī Gumphā monastery.

These, like other sculptures here, are in a poor state of preservation, but they represent the finest achievements at the site. Most remarkable is a long frieze, stretching between the arched doorways of the top story, representing a series of incidents that have not yet been identified. The work parallels that of the Great Stūpa at Sānchi, with the same supple modelling and crowded compositions. At the same time there is a nervous agitation, a fluid, agile movement together with a decided preference for tall, slender human figures. The reliefs on the guard rooms of Rānī Gumphā are also quite remarkable, depicting forested landscapes filled with rocks from which waterfalls flow into lakes that are the sporting grounds of wild elephants. The fine work of this cave strikes a romantic and lyrical note seldom found in Indian art.

Indian sculpture in the 2nd and 1st centuries BC: sculpture in the round and terra-cotta. The most important sculpture in the round are the life-size or colossal images of *yakṣa*s and *yakṣī*s, which reinterpret forms established by the two Patna *yakṣa*s and the Dīdarganj *yakṣī* of the Maurya period—very much as a few animal capitals, particularly the *makaras* (a crocodile-like creature) from Kauśāmbī and Vidiśā (Besnagar), echo the tradition of the superb Maurya animal capitals. It is the *yakṣa* figures, however, that deserve special attention, for they played a significant part in the iconographic developments of the 1st century AD and later and contributed substantially to the imagery of the anthropomorphic Buddha icon.

The most famous of the *yakṣa* images is a colossal figure recovered from the village of Pārkham, near Mathurā (Archaeological Museum). It is about $8^2/_3$ feet (2.6 metres) in height, and, though the two hands are broken and the head is considerably damaged, it is an image of great strength. Its squat neck, its head set close to the body, which tends toward corpulence, its swelling belly restrained by a flat band, and a broad chest adorned with necklaces—all of these features contribute to an image turgid with earthy power. The back is flat and cursively finished, so that the figure has the appearance more of a bifacial relief than of an image carved in the round. Although the forms retain some of the cubical modelling of Bhārhut, the swelling limbs and torso have a massive weightiness that makes the image an appropriate representation of a divinity that presides over the productive processes of nature and endows plenty and abundance on his worshippers.

Mathurā *yakṣa* images

The Mathurā region seems to have been an important centre of *yakṣa* worship, for several images, most of them fragmentary, have been discovered there. Some images have also been found from the ancient city of Vidiśā (Vidisha Museum), one of which is even larger than the Pārkham example and is in a better state of preservation. The god holds a bag in one hand (the other was held below the chest), and the hair is tied in a large top knot over the forehead. The image is accompanied by a female consort (*yakṣī*), wide-hipped and full-breasted, who also emphasizes and personifies the powers of fertility.

The widespread nature of the cult is evidenced by the occurrence of *yakṣa* images throughout India. Fragments in the round (not to speak of the relief representations in a Buddhist context) of the 2nd to 1st centuries BC have been found from Madhyadeśa, Orissa, Rājasthān, Andhradeśa, and Mahārāshtra. At Pītalkhorā there is an exceptionally fine image of a *yakṣa* conceived as a potbellied dwarf carrying a shallow bowl on his head; the features, with a gently laughing mouth, are suffused with good humour. Similar *yakṣa*s, employed as atlantes (male figures used as supporting elements), are also found on the western gateway of the Great Stūpa at Sānchi and at other sites, notably Sārnāth.

The latest in the series of cult images is the image of the Yakṣa Maṇibhadra, from Pawāyā (Gwalior Museum). The sculpture is at present headless, but the rest of the body is well preserved. The right hand holds a fly whisk that flares over the shoulder; the modelling of the legs and torso is sensitive, and the folds of the garment wrapped around the body are full and voluminous, recalling the style of sculpture at Sānchi.

The terra-cotta sculpture of the period consists mainly of relief plaques made from molds found at numerous sites in northern India. These generally depict popular divinities; a richly dressed female figure loaded with profuse jewelry, obviously a mother goddess, is the favoured subject. Scenes from daily life also abound—as well as what appear to be illustrations of current myths and stories. Superb examples have been found from Mathurā, Ahichhatrā, Kauśāmbī, Tāmlūk, and Chandraketugarh. The workmanship is often of the most exquisite clarity and delicacy, the style paralleling that of contemporary stone sculpture.

Indian sculpture from the 1st to 4th centuries AD. This period is characterized by the dominance in northern India of the ancient school of Mathurā. Other schools, such as those that flourished at Sārnāth and Sānchi in the first two centuries before Christ, for example, were markedly restricted in their artistic output. Much of their sculpture was imported from Mathurā, and the few images they produced locally were strongly influenced by Mathurā work. The narrative bas-relief tradition, consisting of elaborate compositions of edificatory character, was on the wane, and the emphasis was on carving individual figures, either in high relief or in the round. For the first time, images appear of the Buddha, *bodhisattva*s, and various other divinities including specifically Hindu images representing the gods Vishnu, Śiva, Varāha, and Devī slaying the buffalo demon; some of these figures begin to feature several arms, a characteristic of later iconography. There are also many images of *yakṣī*s, often in most alluring attitudes and gestures. Their enticing bodies are now presented as unified organic entities, lacking all traces of the stiff, puppet-like aspect that had not been entirely overcome even at the Great Stūpa of Sānchi. During this period, also, a

P Chandra

A Mathurā image of the Buddha discovered at Sārnāth, Uttar Pradesh, India, red sandstone, *c.* AD 81. In the Sārnāth Museum.

fresh incursion of foreign influence by way of western Asia was received, quickly assimilated, and transformed in the characteristic manner of Indian art.

The school of Gandhāra

The school of Gandhāra, with Taxila in Pakistan as its centre and stretching into eastern Afghanistan, flourished alongside the Kushān school of Mathurā. It is of a startlingly different aspect, stressing a relatively naturalistic rendering of form, ultimately of Greco-Roman origin. The school evolved a distinct type of Buddha image and was also rich in relief sculptures depicting Buddhist myth and legend. Drawing largely on Indian traditions of composition, it nevertheless reinterpreted them in its own manner. The schools of Mathurā and Gandhāra were in close proximity and undoubtedly influenced each other, but essentially each adheres to its own concept of style.

The ancient Indian relief style found its fullest expression and development at neither Mathurā nor Gandhāra but in Andhradeśa, notably at the great sites of Amarāvatī and Nāgārjunikoṇḍa. Railing pillars and other parts of *stūpa*s decorated with *Jātaka* tales and scenes from the Buddha's life are found in great number and are of the most exquisite quality. Free-standing images of the Buddha, on the other hand, are relatively rare, being found only toward the close of the period.

Indian sculpture from the 1st to 4th centuries AD: Mathurā. One of the most important contributions of the school of Mathurā was the development of the cult image of the Buddha, who had been previously represented by aniconic (not made as a likeness) symbols. There is a certain amount of controversy about whether Mathurā or Gandhāra originated the Buddha image, which appears to be insoluble in view of the circumstantial nature of the evidence. It is possible that the two schools independently developed their own separate types of images; but, at least as far as the Mathurā image is concerned, it is clear that it is a natural development from the tradition of large *yakṣa* sculptures found in this region. The development can easily be seen in a famous image (discovered at Sārnāth and now in the Sārnāth Museum) of Mathurā manufactured and dedicated by the monk Bala. Carved in the round, the image is shown in a pose of strict frontality, the left hand held at the waist and the right arm, now damaged, originally raised to the shoulder—a posture immediately recalling that of the *yakṣa* images. The jewelry, however, is appropriately omitted, and the body is clothed in simple monastic garments. The modelling throughout is strong and sensuous, and the radiant energy of the body, its affirmative, outgoing movement, is more appropriate to the personality of a *yakṣa* than to that of the Buddha. This standing Buddha image, as seen in the Bala statue, is the standard Mathurā type, several examples of which are known. Along with this one, a similar, seated type developed, of which the best example is the splendid image known as the Kaṭra Buddha (Archaeological Museum). The modelling of the body is refined, the breasts characteristically heavy and prominent, and the flesh of the torso, with its subtle modulations, as convincingly rendered as the Bala image.

Stone tablets

The new trends formulated early by the Mathurā school do not indicate a sharp break from the traditions of the earlier schools. This is clear in a series of magnificent *āyāgapaṭas,* or stone tablets originally set up outside *stūpa*s to receive worship and offerings. They are usually square or rectangular and richly decorated with auspicious and religious symbols as well as angelic and mythical beings. The extremely decorative, lavish surface treatment gives the immediate impression of a great profusion of multiple forms, akin in feeling to the sculpture of the Great Stūpa of Sānchi. The organization of these forms, however, has none of the easy freedom of Sānchi. The figures, for example, are often cast in a regular, winding shape imitating the movement of the undulating lotus creeper. The same movement is seen in rows of animals depicted with haunches raised and chests touching the ground, features seen in earlier art but now much more emphatically stylized. The bodies of the animals also begin to be overpowered by vegetal forms, the tails, for example, terminating in foliate tips; in a later age, this tendency results in the almost total disintegration of animal shapes under the pressure of the floral.

It is not to these bas-reliefs, however, that one turns for the most delightful creations of the Mathurā school (for they are in fact the last vestiges of a style rapidly passing out of favour) but to the large number of railing pillars usually carved with representation of *yakṣī*s engaged in playful and enticing activities such as plucking blossoms from trees or leaning on its branches, dancing, bathing under a waterfall, and adorning themselves. Among the most beautiful of these is a group that was recovered from Kaṅkālī Ṭīlā and now in the State Museum at Lucknow. The modelling of the figures is generally heavy, the soft, plump bodies suffused with a slow, languorous movement. What is important, however, is the emotion, which is no longer expressed in the face alone but in the whole attitude of the body. The pensive mood of a woman holding a lamp, for instance, is evoked not only by the serene features of the face but by the gentle sway of the relaxed body. Present throughout is a fresh movement of life, a marked striving for diverse and varied effects of posture, movement, expression, and even dress and ornament that brings about vital changes in the nature of Indian sculpture. A remarkable group of railing posts decorated with *yakṣī* images, which were recovered from Bhūtēsar near Mathurā (Archaeological Museum), represent an even more refined achievement than the Kaṅkālī Ṭīlā figures. The heavy proportions, in spite of the full breasts and the wide hips, have been overcome; the happy faces express carefree joy, and the postures of the body are so alive with rhythm as to give the impression of a dancing figure.

Mathurā, during this period, was ruled by the Kushān (Kuṣāṇa) dynasty. A group of portrait sculptures of these rulers (Archaeological Museum), recovered from a village called Māt in the environs of Mathurā, gives an interesting glimpse of the foreign influences entering India at the time. One of them (unfortunately lacking the head) represents the emperor Kaniṣka wearing heavy boots, a tunic, and a coat, and leaning on a mace. The image is quite

P. Chandra

Headless portrait statue of the emperor Kaniṣka from Mathurā, Uttar Pradesh, India, *c.* late 1st century AD. In the Archaeological Museum, Mathurā, Uttar Pradesh, India. Height 1.85 m.

different not only in dress but also in style from other contemporary works, being essentially linear, with the forms entirely set into the surface. The surfaces have little ornamentation and are marked by extreme simplicity; they are also uncompromisingly stiff and rigid. It is possible that these images represent attempts by a Mathurā artist to imitate a style preferred by his imperial masters; but it was not long before the foreign elements were assimilated into the Mathurā style proper, for later images of Kushān chiefs have the same expanding and voluminous form that characterizes other sculptures of this school. A large number of ornamental motifs that now appear in India for the first time undergo a similar process of transformation.

The extent of Mathurā influence on Indian art of this period can be gauged by the sculpture of the school found at several sites in different parts of northern India, notably Ahichhatrā, Kauśāmbī, Sārnāth, and Sānchi. Most of these sites had been flourishing centres earlier, but only a very limited amount of sculpture was produced during the ascendancy of the Mathurā school; and whatever local sculpture was produced at this time was heavily influenced by the Mathurā style. At Sārnāth, for example, both the Bala Buddha imported from Mathurā and its local imitations have been found.

Ivory plaques of Bagrām

Ivory plaques discovered at Bagrām (Begrām) in Afghanistan are closely related to the school of Mathurā. These are of great importance; for, though ivory must have been a favourite medium of sculpture, little has been preserved of the early work. Most of it is in very low engraved relief, with fluent, sweeping outlines. The figures are depicted in easy and elegant postures, and the workmanship often attains considerable virtuosity.

P Chandra

Bodhisattva of the Gandhāra school, schist sculpture, *c.* 2nd century AD. In the Municipal Museum, Allāhābād, Uttar Pradesh, India.

Indian sculpture from the 1st to 4th centuries AD: Gandhāra. Contemporary with the school of Mathurā, and extending almost into the 6th century, is the Gandhāra school, whose style is unlike anything else in Indian art. It flourished in a region known in ancient times as Gandhāra, with its capital at Taxila in the Punjab, and in adjacent areas including the Swāt Valley and eastern Afghanistan. The output of the school was very large; numerous images, mostly of Buddhas and *bodhisattvas*, and narrative reliefs illustrating scenes from the Buddha's life and legends have been found. The favoured material is gray slate or blue schist and, particularly during the later phases, stucco. Except for objects excavated at a few well-known sites (such as Taxila, Peshāwar, and the Swāt Valley, in Pakistan, and Jalālābād, Haḍḍā, and Bāmiān, in Afghanistan), most of the finds have been the result of casual discovery or clandestine treasure hunts and plunder, so that their correct provenance is not known. If to this are added the large variety of idioms that appear to have existed simultaneously and the total absence of securely dated images, the wide divergence of scholarly opinion with regard to the schools' evolution can be understood. In the present day, there is general agreement, however, that its most flourishing period probably coincided with Kushān rule, particularly the reigns of the emperor Kaniṣka and his successors, and that the school did not long outlast the growth of the Gupta school in the 5th century.

The origins of the Gandhāra style are ultimately Greco-Roman, though, recently, emphasis has been placed on Roman art as the more immediate source. It has also been suggested that the school was created by foreign craftsmen imported into India and by their Indian pupils.

The Gandhāra school is also credited by some scholars with the invention of the anthropomorphic Buddha image. Whether this is correct or not, the Gandhāra image is quite different from that of Mathurā and illustrates the difference between the two schools. Instead of the powerful images directly descended from *yakṣa* prototypes, the Gandhāra version is an adaptation of an Apollo figure, with rather sweet and sentimental features. The definite volume and substance given to the pleated folds of the monastic robes make this image more naturalistic than anything found in Indian art. At the same time, the iconographical features are of Indian origin. Large numbers of *bodhisattva* images conceived in the image of royalty, some with strongly individualized facial features, have also been found.

Narrative relief sculpture

In contrast to Mathurā, narrative relief sculpture was very popular in Gandhāra art. Again, in composition and iconography these reliefs are largely dependent on the earlier Indian schools, but the style is quite distinct. Instead of continuous narrative, incidents separated in time are separately represented, though often arranged in sequence. Violent emotions are realistically rendered. The compositions range from simple horizontal placement of figures to rich and complex arrangements, which often attempt to render space illusionistically.

In the course of time, Indian influence was increasingly felt in the art of Gandhāra, and an abstract vision began to obscure the Greco-Roman naturalism of the earlier forms. In spite of the new influence (and the many graceful but cloying stucco sculptures that are representative of this late phase) the style shows no signs of vital change. This conservatism, together with the large artistic production, gives an overall impression of considerable monotony. Without any real roots in India and with marked foreign features, the avenues of natural development seem to have been closed to the school, which thus finally disappeared. Nevertheless it made vital contributions to the art of Central and eastern Asia, and several features, drastically transformed, were incorporated in Gupta art.

Indian sculpture from the 1st to 4th centuries AD: Andhradeśa. Besides the schools of Mathurā and Gandhāra, a most accomplished school of sculpture flourished in Andhradeśa during the three centuries after Christ, the most important centres being Amarāvatī and Nāgārjunakoṇḍa. The remains consist mainly of carved railings and rectangular slabs that decorated the great Buddhist *stūpas*, which have largely disappeared. The finds are thus frag-

mentary and belong to several phases of construction or to separate monuments spanning the 1st, 2nd, and 3rd centuries AD.

Unlike the school of Mathurā, which concentrates on the carving of single figures, the Amarāvatī school carried to the fullest limit of its development the ancient tradition of relief sculpture, which flourished in the two centuries before Christ at sites such as Bhārhut, Sānchi, and Amarāvatī itself. The marble railing posts are decorated with central medallions and lunates at the top and bottom, all filled with lotus flowers of a very rich design. Often the medallions also contain reliefs illustrating scenes from the Buddha's life and from the *Jātaka* stories, and these are the principal glory of the site.

Two broad phases in the development of narrative relief can be distinguished. In the first, the artist builds on the achievements of early relief sculpture as seen on the Great Stūpa of Sānchi. The forms are still comparatively heavy, the figures increasingly soft and fleshy, the movement freer but still pervaded by a sense of calm repose. This type of work, represented by relatively few examples, is followed by a phase in which the compositions achieve an extraordinary elaboration and complexity. Most striking is the restless, energetic movement, often nervous and flurried, that possesses the participants in any given scene. Complex relationships and patterns are established between the figures; and space is so articulated that the eye participates in the swirling inner movement of the composition that effectually dissolves the ground on which the figures are carved, while the figures themselves flow out in an endless movement from the ground. The setting is dramatic in the extreme. The loving workmanship, reminiscent of ivory carving, and the superb technical proficiency mark the Amarāvatī reliefs as the culminating point of the entire relief style.

The figures, of both men and women, are of unprecedented suppleness and plasticity, the forms rendered in every variety of torsion and flexion. A fluent, gliding line, often more appropriate to painting than to sculpture, encloses the figures, and pervading the whole is a subtle voluptuousness. The reliefs are often only nominally religious, a pretext for the sculptor's pleasure in representing the leisured and sophisticated life of the time.

Nāgārjunakonḍa sculpture marks the last phase of the relief style. The figures become stiffer and puppet-like, the patterns of movement frozen and mechanical but still possessing the energy and richness that always characterize this style.

Buddha images of Andhradeśa

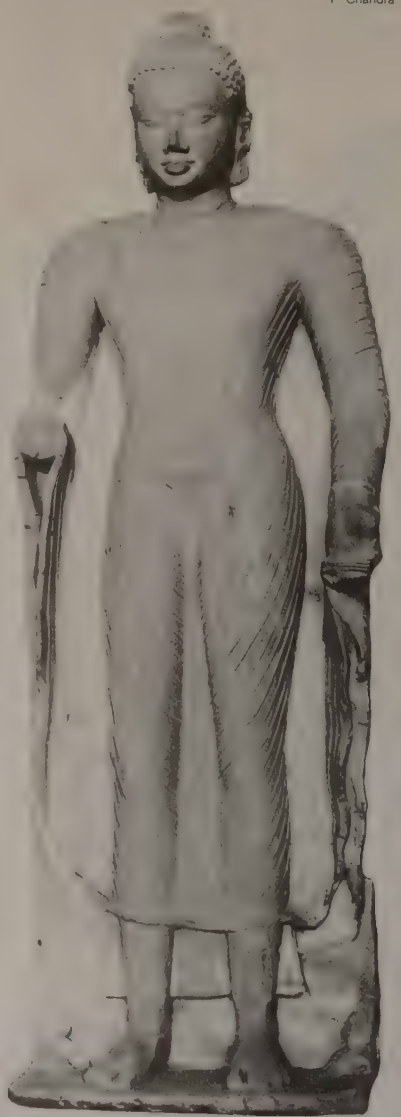
The Buddha is represented in Andhradeśa by both symbolic and anthropomorphic forms. The iconographic formula developed shows him clad in a rather thick garment with stylized folds, and the postures are not as formal and hieratic as the Mathurā. This type of Buddha exercised considerable influence in the development of the Buddha image in Ceylon. In several other features as well, the Andhra style also contributed to the development of early sculpture in Southeast Asia.

Indian sculpture from the 1st to 4th centuries AD: terracotta. The quality of terra-cotta figurines of this period is generally inferior to work produced in the first two centuries BC. Many heads of crude workmanship, with protruding eyes, apparently representing foreigners, were found at sites such as Mathurā, Ahichhatrā, and Kauśāmbī. At the same time, there are some well-modelled heads that imitate the style of stone sculpture and are equally expressive.

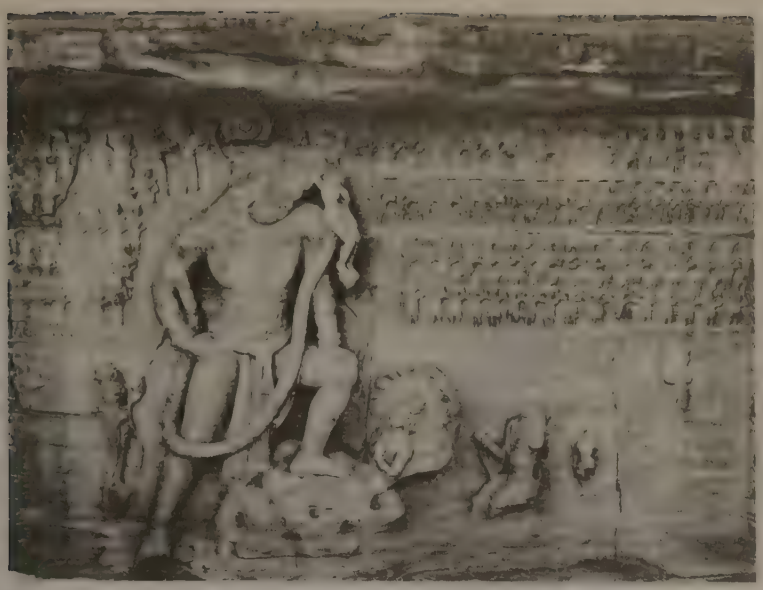
Gupta period (c. 4th–6th centuries AD). During the 4th and the 5th centuries, when much of northern India was ruled by the Gupta dynasty, Indian sculpture entered what has been called its classic phase. The promise of the earlier schools was now fully realized, and at the same time new forms and artistic ideals were formulated that served as the source for development in succeeding centuries. The more or less sensuous and earthy rendering of form was drastically transformed, so that artistic expression closely conformed to the religious vision. The forms are refined and treated with sure and unsurpassed elegance. The volumes, impelled by an inner life, still swell from within but are restrained and controlled, made to flow in smooth and abstract rhythms in an organic and unified concept in which the sensual and the spiritual are inextricably blended. The edificatory, didactic intent of early relief sculpture is abandoned; instead, the works produced are pronouncedly meditative; and the repose and calm that settles on the images of the Buddha, the master of the inner contemplative life, is also seen on images of other divinities. Decorative ornament is in perfect harmony with the volumes it adorns, each emphasizing the other, so that in every respect this classic style of the Gupta period is one of great composure and perfect balance.

Gupta period: Mathurā. The impetus for the new schools seems to have come from Mathurā, which is hardly surprising in view of the preponderant role played by the city in the preceding period. The transformation into the new idiom is best illustrated by a splendid image of the Buddha which is dated AD 384 (Indian Museum, Calcutta). Memories of the rather massive and ponderous weight of the earlier style are present, but the calm face no longer looks out at the world; rather, the vision is turned within, the mood being one of serene contemplation. The style, which consistently uses the local red sandstone, undergoes further refinement, seen in a series of magnificent life-size Buddha images of the 5th century (now scattered in museums throughout the world). The more delicate face radiates a feeling of calm inner bliss, and the body is most subtly modelled by smoothly flowing planes that both suggest the swelling force of life and subordinate it to

P Chandra



Buddha, red sandstone sculpture from Mathurā, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.



Vishnu rescuing the Earth goddess, sandstone relief panel from a cave at Udayagiri, Madhya Pradesh, India, early 5th century AD.
P. Chandra

the spiritual vision of the whole. Mathurā images generally show the Buddha wearing a diaphanous robe, the folds of which are rendered by stringlike ridges in a reinterpretation of a Gandhāra convention. The gestures of the hand are delicate and varied. The hair is usually rendered by rows of small curls that conceal the conical protuberance. These Mathurā images established an iconographical type that became the norm for the Buddha image.

P. Chandra

Buddha, Chunār sandstone sculpture from Sārnāth, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.

Mathurā Hindu images

In addition to the Buddha figure, Mathurā has yielded large numbers of images of the various Hindu divinities, particularly Vishnu-Krishna. This is in keeping with the increasing strength of the various Hindu cults and the intimate association of Mathurā with the god Krishna. The famous image of Vishnu from Kaṭrā Keśavadeva in Mathurā is one of the finest (National Museum, New Delhi). The god is conceived as a royal figure, wearing a crown and appropriate jewelry, his features imbued with a dignified calm that is suitable to his function as the preserver and is also characteristic of most Gupta art.

Gupta period: Sārnāth. This famous centre of Indian art developed a sweeter and more elegant version of the Buddha image than Mathurā's. Instead of the rather strict frontal posture, the weight of the body is thrown more on one leg, resulting in a very subtle contrapposto position, in which the hips, shoulders, and head are turned in different directions. This lends a certain movement to the figure, so that it does not quite possess the static, steadfast quality of Mathurā. The robes are no longer ridged with folds but are plain, and the surface of the stone is even more abstractly handled than is the Mathurā. The faces are heart-shaped, the transitions from one part of the body to another smoother, so that the images have great refinement even if they do not possess the strength of Mathurā. The characteristic Sārnāth style, the preferred material of which is the local buff Chunār sandstone, seems to have developed in the late 5th century, the few earlier works being closer to the Mathurā school. The most famous image from the site and one of the masterpieces of Indian art is that of the seated Buddha preaching (Sārnāth Museum). It is exceptionally well preserved and delicately carved. The face, with serene features and a gentle smile playing on the lips, suggests the joy of supreme spiritual achievement. The halo behind the Buddha is also very beautifully carved, with exquisite floral patterns. Large numbers of Buddha and *bodhisattva* images have been excavated at Sārnāth and are to be found in the museum at the site and in major collections throughout the world.

Gupta period: central India. In addition to the major schools of Sārnāth and Mathurā, important sculpture of the 5th and 6th centuries is found at several sites in central India. The sculptures here are often in their original locations, surviving not as isolated images torn from their architectural context but in association with the temples of which they formed a part. At Udayagiri, near Vidiśā, are a series of simple rock-cut caves of the opening years of the 5th century. The sculpture, made of soft stone, has suffered greatly, but whatever has survived reveals a style that stresses strength and power. Perhaps the most

magnificent work is a great relief panel depicting the boar incarnation of Vishnu lifting the earth goddess from the watery deeps into which she had been dragged by a demon. The massive figure of the god, with the body of a man and the head of a boar, is carved in a surging movement across the face of the rock, the goddess resting easily on his shoulder, while a host of beings, human and divine, celebrate this great triumph.

The Śiva temple at Bhumarā has also yielded some sculpture of fine quality. The stone is carved with great precision and skill, nowhere more evident than in the handling of exuberant floral ornament. Little in Indian decorative sculpture can match the brilliance of the large panels filled with lotus stems and floriated scrolls discovered at this site and at Nāchnā Kutharā.

Some of the finest Gupta sculpture adorns the walls of the Vishnu temple at Deogarh. Particularly striking are three large relief panels depicting Vishnu lying on the serpent Śeṣa, the elephant's rescue, and the penance of Nara-Nārāyaṇa. The compositions tend to be dramatic; the carving and decoration, sumptuous, the sturdy forms recalling Mathurā rather than the attenuated grace of Sārnāth. The doorframe of the sanctum of this temple is an especially fine example of architectural decoration popular in this period. Bands of floral scrolls, amorous couples, and flying angels of great elegance are carved around the entrance. Particularly impressive are groups of worshippers at the base, their swaying bodies related to each other with an easy rhythm.

Gupta period: Mahārāshtra. A great revival of artistic activity seems to have taken place in this region during the reign of the Vākāṭaka dynasty and its successors, best expressed in the splendid sculpture decorating the cave temples of Ajantā and Elephanta. The idioms established in the North were adapted here to the needs of a style that conceived figures on a massive scale, as determined by the demands of the great expanses of rock out of which they were carved. Although the sculpture at Ajantā (mostly of the late 5th century) combines the old weightiness with the new restraint and elegance, the style finds its supreme expression in the magnificent cave temple at Elephanta. The central image of this great temple is of immense size and in deep relief. It represents Śiva in his cosmic aspect, the central head clam, introspective, self-sufficient, and transcending time, the heads to the sides, in their sensuous beauty and awesome terror, reflecting the creative and the destructive aspects of the supreme divinity.

Sculpture of the Ajantā and Elephanta caves

Gupta period: other regions. The impact of the Gupta style of the 5th and 6th centuries was felt in many parts of India, though actual remains thus far discovered are more abundant in some parts than in others. There appears to have been, in Bihār, a distinct school characterized by rather heavy, compact forms; and Gujarāt and southern Rājasthān developed an individual style of considerable voluptuousness and plasticity. Among the notable sculpture of the Idar region are groups of mother goddesses whose massive forms are rendered with an easy grace and intimacy. In the Karnataka country, to the south, the cave temples of Bādāmi reveal yet another distinct idiom, somewhat direct and elemental but nevertheless belonging to the same general style, with local variations, that prevailed over the greater part of India.

Gupta period: terra-cotta. Terra-cotta sculpture, like art in other mediums, was greatly developed. Fairly large and elaborate plaques were used to adorn brick *stūpa*s and Hindu temples from Sind to Bengal. The polychrome relief images of the Buddha from Mīrpur Khās are delicate and slender, with traces of Gandhāra feeling. Representations of divinities and mythological scenes from temples in Bīkaner, Ahichhatrā, Bhītargaon, and Śrāvastī are works on a more popular level, possessing an earthy ponderousness. A large number of figurines, particularly fragments of heads with elaborate coiffures and delicate, smiling features, have been found at Rājghāt in Vārānasi (Benares) and at other sites.

Medieval Indian sculpture. Indian sculpture from the 7th century onward developed, broadly speaking, into two styles that flourished in northern and southern India, respectively. In each of these regions there also developed additional local idioms, so that there was a wide variety of schools. All, however, evolved in a consistent manner, the earlier phase marked by relatively plastic forms, the later phase by a style that emphasizes a more linear rendering. The sculpture was used mainly as a part of the architectural decor, and the quantity required was vast. This often entailed a mechanical production, with the result that works of quality are few in proportion to the numbers.

Detail of a wall of the Lakṣmaṇa temple at Khajurāho, Madhya Pradesh, India, sandstone, *c.* AD 941.
P Chandra

Besides the two main idioms, the local schools of Mahārāshtra and Karnataka are of particular interest because they possess considerable individuality and often show both northern and southern features.

Sculpture in bronze was also produced in fairly large quantities in this period. Again, several local schools can be distinguished, the most important of which are those of eastern and southern India.

Medieval Indian sculpture: North India. The history of North Indian sculpture from the 7th to the 9th centuries is one of the more obscure periods in Indian art. Two trends, however, are clear: one exhibits the decline and disintegration of classical forms established during the 5th and 6th centuries; and the other, the evolution of new styles that began to possess overall unity and stability only in the 10th century.

A breakdown of the Gupta formula is observable from at least the 7th century onward, if not a little earlier: harmonious proportion, graceful movement, and supple modelling begin to yield to squat proportions, a halting movement, and a more congealed form. Toward the 8th century, signs of a new movement become evident in a group of sculptures that departs from the progressively lifeless working out of the Gupta idiom. The modelling emphasizes breadth but with a pronounced feeling for rhythm, and the delineation of decorative detail is fairly restrained. In the 9th century, particularly during the second half, a distinct change came over the styles of all of northern India. A new elegance, a richer decorativeness, and a staccato rhythm so characteristic of the medieval styles of the 10th and 11th centuries begin to be clearly seen and felt. Sculpture of this period reaches a standard of elegance never surpassed in the medieval period: the grace and voluminousness of earlier work are modified

Signs of a new movement

but not lost; the harsh angularity of later work, avoided. An idea of the style can be formed from an important group of sculptures at Ābānerī, the Śiva temple at Indore, and the Telī-kā-Mandir temple at Gwalior, as well as from individual works in various North Indian museums.

With the 10th century, the conventions of North Indian sculpture became fairly well established. The style is represented by examples from such monuments as the Lakṣmaṇa temple at Khajurāho (dated 941), the Harasnāth temple at Mt. Harsha (*c.* mid-10th century), in Rājasthān, and numerous other sites scattered all over northern India. These works are executed in a style that has become harder and more angular, the figures covered with a profusion of jewelry that tends to obscure the forms it decorates. These features are further accentuated in the 11th century, when many temples of great size, adorned with prodigious amounts of sculpture, were erected all over northern India. There is a decline in the general level of workmanship: the carving is often entirely conventional and lifeless, the features rigid and masklike, and the contours stiff and unyielding. The ornamentation, consisting of a profusion of beaded jewelry, is for the most part as dull, repetitive, and lifeless as the rest of the sculpture. This phase of artistic activity is represented at important centres from Gujarāt to Orissa; one of them is Khajurāho, with a vast amount of sculpture, all in a good state of preservation but conceived and executed as perfunctory architectural ornamentation. Not all sculpture, however, is of inferior quality; the hard, metallic carving and angular, stylized line sometimes result in works possessing a cold brilliance.

The 12th century marks the end of traditional sculpture all over northern India, except for a few pockets not yet penetrated by the Islāmic invasions. A rigid line imposed itself on the forms, which in turn became desiccated and hard, so that whatever unity of surface may have existed was entirely shattered. A brief revival took place in parts of Gujarāt and Rājasthān in the 15th century, but the sculpture merely imitated the work of the late medieval phase. The pure geometry of their forms, however, sometimes results in works possessing a curious archaistic power.

Sculpture in eastern India

Sculpture in eastern India (consisting of Bangladesh and the modern Indian states of Bihār, West Bengal, and Orissa), though sharing in the broad pattern of development of the rest of northern India, nevertheless represents a distinct idiom. The flatness of planes and angularity of contours are less pronounced, the figures retaining a sense of mass and weight for a greater period of time and to a greater degree. This can be clearly seen in sculpture from Konārak, in Orissa. Dating to the 13th century, the style retains a considerable semblance of plasticity at a period when sculpture in other parts of northern India had assumed a very wooden appearance. In Bihār and Bengal a flourishing school of bronze sculpture also developed, as evidenced by the large number of finds, notably from the sites of Nālandā and Kurkihār. The style generally parallels works in stone, emphasizing plastic values to a great degree. The most flourishing period was the 9th century, when a series of magnificent images representing the gods and goddesses of the Buddhist pantheon were made at Kurkihār and Nālandā. The work of the 10th and 11th centuries is more decorative and often very skillfully and elaborately cast. Of relatively small size and therefore easily transportable, bronze sculpture from this area played an important part in the diffusion of Indian influence in Southeast Asia.

Kashmir sculpture tends to be weightier and more massive than works in other parts of India. Some Gandhāra memories surivive, particularly in the fleshy rendering of the body and the drapery, but the sculpture is very much a part of the stylistic developments in northern India. Representative examples of the style, dating to around the mid-9th century, have been found from Avantipura. A flourishing school of bronze sculpture also existed, numerous examples having come to light in recent years. One of the finest, discovered at Devsar (Sir Pratap Singh Museum, Srinagar), is a large 9th-century ornamental frame, 6½ feet (two metres) high, decorated with various incarnations of Vishnu, all filled with great energy and movement. A good number of ivory images of Kashmir

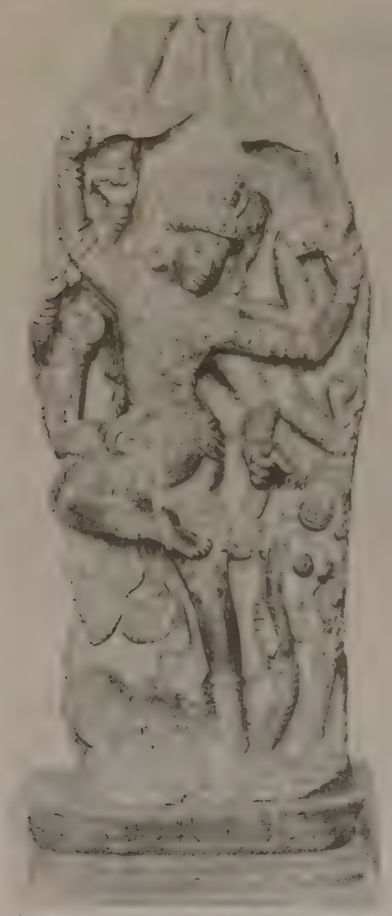
Seated Buddha with attendants, carved ivory sculpture from Kashmir, *c.* 8th century AD. In the Prince of Wales Museum of Western India, Bombay. Height 10 cm.
P Chandra

workmanship have also been preserved. These are generally of miniature size, polychromed, and of extremely fine and delicate workmanship. Influences of the Kashmir style of sculpture were strongly felt in the neighbouring Himalayan region, including both Tibet and Nepal.

Medieval Indian sculptures: southern India. The medieval phase in southern India opened with elegant 7th-

P Chandra

Śiva-Ardhanārīśvara, granite sculpture from the Ugrakaliamman temple, Thanjāvūr, Tamil Nadu, India. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height 1.14 m.



Śiva slaying the elephant demon, granite sculpture from Dārāsuram, Tamil Nadu, India, 13th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu.
P Chandra

century sculptures at Mahābalipuram, by far the most impressive of which is a large relief depicting the penance of Arjuna (previously identified as an illustration of the mythical descent of the Ganges). It is carved on the face of a granite boulder with a deep cleft in the centre, representing a river, down which water actually flowed from a reservoir situated above. On both sides are carved numerous figures of divinities, human beings, and animals that crowd the hermitage where Arjuna, practicing penance, is visited by Śiva. The tall, slender figures, with supple tubular limbs, remotely recall the proportions of Amarāvatī, now greatly transformed; and the numerous animals, including the elephant herd with its young, show the same intimate feeling for animal life that characterizes all Indian sculture, but in a manner that has seldom been surpassed.

The light, aerial forms gained stability and strength in subsequent centuries, culminating in superb sculptures adorning small, elegant shrines built during the late 9th century when the Cōḷa dynasty was consolidating its power. The temples at Tiruvalīśvaram, Kodumbālūr, Kilaiyur, Śrīnivāsanālūr, Kumbakonam, and a host of other sites of this period are only sparingly adorned with sculpture, but it is of superb quality. With the 10th and 11th centuries, South Indian sculpture, like its counterpart in the north though to a lesser degree, was carved in flatter planes and more angular forms, and the fresh, blooming life of earlier work is gradually lost. This can be seen, for example, in the sculpture of the numerous temples of Thanjāvūr and Gaṅgaikoṇḍacōḷapuram. The subsequent phase, extending up to the 13th century, is represented by work at Dārāsuram and Tribhuvanam; although the forms become increasingly congealed, brittle works of fine quality—often capturing outer movement with great skill—continue to be produced. Sculpture in southern India continued when artistic activity was interrupted in the north by the Islāmic invasions but, in spite of technical virtuosity, became progressively lifeless. Artistic activity continued in the south into the 17th century, the elaborately sculptured halls at Madurā and the masses of stucco sculpture adorning the immense entrances, or *gopuras*, testifying to the prodigious output and the undistinguished quality of the work produced.

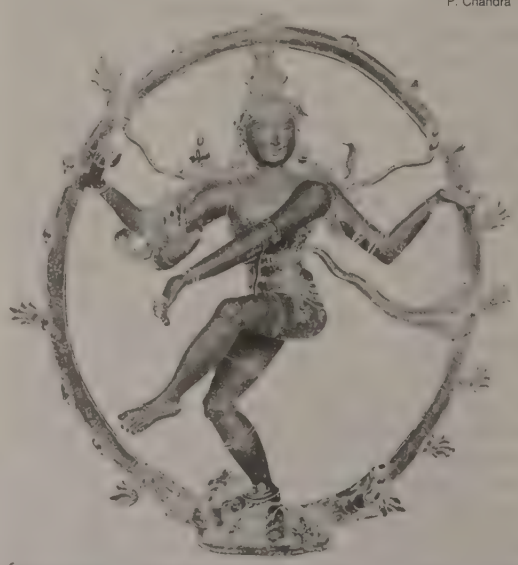
South Indian bronze sculpture

South Indian bronze sculpture has a special place in the history of Indian art. A large number of images were made (some of them still in worship in the mid-20th century and others unearthed from the ground by chance), but examples before the 8th century are quite rare. In bronze, as in stone, the 9th and 10th centuries were periods of high achievement, and many images of excellent quality have survived. They are all cast by the lost-wax, or cire perdue, process (in which a wax model is used) and technically are very accomplished. In the early stages the forms were smooth and flowing, with a fine balance maintained between the body and the complex jewelry, the lines of which follow and reinforce every movement of the plastic surface. The bronzes of the later period lose this cohesiveness, the ornament, by virtue of its hardness, tending to divide and fragment the body it covers. The modelling also became flatter and sharper, though not quite as rapidly in bronze sculpture as in stone. Ancient traditions of workmanship survive to the present day, and a few guilds of craftsmen continue to make competent if somewhat lifeless images.

Most South Indian bronze images are representations of Hindu divinities, notably Vishnu and Śiva. One particular form deserves special notice as a striking southern contribution to Indian iconography. It is that of a four-armed Śiva as Lord of the Dance (Naṭarāja), shown within a flaming halo, or aureole, one hand holding the doubleheaded drum symbolizing sound, or creation, and the other holding the fire that puts an end to all that is created. The palm of the third hand faces the devotee, assuring him of freedom from fear, while the fourth hand points to the raised foot, the place of refuge from ignorance and delusion, which are symbolized by the dwarf demon crushed beneath the other foot. Several splendid images are known, the finest being, perhaps, the great image still worshipped in the Bṛhadīśvara temple at Thanjāvūr.

Medieval Indian sculpture: Mahārāshtra and Karnataka. The Karnataka country possessed a flourishing school of sculpture in the 7th and 8th centuries, as seen in examples from Aihole, Pattadkal, and Ālampur. As in architecture, influences from the north are discernible, but the style is basically southern, emphasizing rugged strength and power compared to the more elegant and delicate forms of the Tamil country. In Mahārāshtra, cave temples at

P. Chandra



Śiva-Naṭarāja, bronze sculpture from Kivalur, Tamil Nadu, India, 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height 81 cm.

Dancing Kālī, relief from the Rameśvara cave, Ellora, Mahārāshtra, India, *c.* 7th century AD.
P. Chandra

Ellora carry the most important examples of this phase of sculpture. Here the tradition is continued of images of great size that, in their primitive strength, partake of the nature of the rock out of which they are carved. A series of large, splendid panels (6th century AD) depicting incidents from Hindu mythology in high relief are to be found in the Rameśvara cave; notable among them is a fearsome representation of the dancing Kālī, goddess of death. The Kailāsa temple (*c.* 757–783) has a remarkable group of elephants struggling with lions all around the plinth. Of the several large reliefs, also at Kailāsa, the depiction of Rāvana shaking Kailāsa is a composition of considerable grace and charm.

Distinctive style of Karnataka

Toward the 13th and 14th centuries, a very distinctive style developed in the Karnataka country, which was then largely ruled by kings of the Hoysaḷa dynasty. The materials employed are varieties of stone that are soft when freshly quarried but harden on exposure, which may account partially for the extreme richness of the work. The sculpture is in very high relief, often undercut and literally covered with the most elaborate ornaments and jewelry from top to toe. This unrestrained extravaganza is unique even for Indian art, which shows a preference for intricate and elaborate ornament at all stages of its history.

Painting. Literary works testify to the eminence of painting as an art form in India, particularly in the decoration of walls, but climate has taken a devastating toll, leaving behind only a few tantalizing examples. By far the bulk of the preserved material consists of miniature painting, initially done on palm leaf but later on paper. The subject matter is generally religious (illustrating divinities, myths, and legends) and literary (illustrating poetry and romances, for example), though the Mughal school was also concerned with historical and secular themes. The styles were rich and varied, often closely connected with one another and sometimes developing and changing rapidly, particularly from the 16th century onward. The work also shows a surprising vitality under strained circumstances, surviving up to the very eve of the modern period when the other arts had deteriorated greatly.

Prehistoric and protohistoric periods. Painting in India should have a history stretching as far back as any of the other arts but, because of its perishable nature, little has survived. None of the examples found in rock shelters over almost all of India, and chiefly representing scenes of hunting and war, appears to be earlier than the 8th century BC, and all may be as late as the 10th century AD. A faint idea of the painter's art in the Indus Valley civilization can be had from the pottery, elaborately decorated with leaf designs and geometrical patterns.

Paintings at Ajantā

Ancient wall painting. The earliest substantial remains are those found in rock-cut cave temples at Ajantā, in western India. They belong to the 2nd or 1st century BC and are in a style reminiscent of the relief sculpture at Sānchi. Also found at Ajantā are the most substantial remains of Indian painting of about the 5th century AD and a little later, when ancient Indian civilization was in full flower. The paintings, the work of several ateliers, decorate the walls and ceilings of the numerous cave temples and monasteries at the site. They are executed in the tempera technique on smooth surfaces, prepared by application of plaster. The themes, nominally Buddhist, illustrate the major events of the Buddha's life, the *Jātaka* tales, and the various divinities of the expanding Buddhist pantheon. The ceilings are covered with rich motifs, based generally upon the lotus stem and the world of animals and birds. The style is unlike anything seen in later Indian art, expansive, free, and dynamic. The graceful figures are painted by a sweeping and accomplished brush; and they are given body and substance by modelling in colour and by a schematic distribution of light and shade that has little to do with scientific chiaroscuro. The narrative compositions, handled with utmost dexterity, are a natural outgrowth of the long traditions of relief sculpture and reflect the splendour and maturity of contemporary sculpture. The large images of the *bodhisattva*s in Cave 1, combining rich elegance with spiritual serenity, reflect a vision that sees the shifting world of matter and the transcendental calm of Nirvāṇa as essentially one.

Except for a large and magnificent painting of a dance scene found at the rock-cut cave at Bāgh—a painting executed in a style closely resembling Ajantā—hardly any other work of this great period survives. Cave temples at Bādāmi, in the Karnataka country, and Sittānavāsal, in Tamil Nadu, probably of the late 6th and 7th centuries AD are already but echoes of the style of the 5th century, which appears to have died out around this time.

Eastern Indian style. Small illustrations on palm leaf, chiefly painted at the great Buddhist establishments of eastern India, appear to have conserved some elements of this ancient style; but they have lost its dramatic impact, which is replaced by a studied preciosity and an inhibited meticulousness. The surviving paintings date from the 11th and 12th centuries and are conventional icons of the numerous Buddhist gods and goddesses, narrative representations having largely disappeared. With the destruction of these Buddhist centres by the Islāmic invader, the east Indian style seems to have come to an end.

Western Indian style. The style of Ajantā is succeeded in western India by what has been appropriately named the western Indian style. Among the earliest examples are a few surviving wall paintings of the Kailāsa temple (mid-8th century) at Ellora and the Jaina temples, built at the same site a hundred years later. The plastic sense of form, so evident at Ajantā, is emphatically replaced by a style that even at this early stage is heavily dependent on line. The contours of the figures are sharp and angular, the forms dry and abstract; and the fluent, stately rhythms of Ajantā have become laboured and halting.

Jaina manuscripts

The most copious examples of this style, however, have survived not on the walls of temples but in the large number of illustrated manuscripts commissioned by members of the Jaina community. The earliest of these are contemporary with eastern Indian manuscripts and are also painted on palm leaf; but the style, instead of attempting to cling to ancient traditions, moves steadily in the direction already established at Ellora. It is a perfect counterpart of contemporary sculpture in western India, relying for its effect on line, which progressively becomes more angular and wiry until all naturalism has been deliberately erased. The figures are almost always shown in profile,

the full-face view generally reserved for representations of the *tīrthaṅkaras*, or the Jaina saviours. A convention that appears unfailingly for the duration of the western Indian style is the eye projecting beyond the face shown in profile, meant to represent the second eye, which would not be visible in this posture. The colours are few and pure: yellow, green, blue, black, and red, which was preferred for the background. In the beginning, the illustrations are simple icons in small panels; but gradually they become more elaborate, with scenes from the lives of the various Jaina saviours as told in the *Kalpa-sūtra* and from the adventures of the monk Kālaka as related in the *Kālakācāryakathā* the most favoured.

Even greater elaboration was possible with the increasing availability of paper from the late 14th century; with larger surfaces to paint on, by the middle of the 15th century artists were producing opulent manuscripts, such as the *Kalpa-sūtra* in the Devasanopaḍā library, Ahmadābād. The text is written in gold on coloured ground, the margins gorgeously illuminated with richest decorative and figural patterns, and the main paintings often occupying the entire page. Blue and gold, in addition to red, are used with increasing lavishness, testifying to the prosperity of the patron. The use of such costly materials, however, did not necessarily produce works of quality, and one is often left with the impression of cursive and hasty workmanship. With some variations—but hardly any substantial departures from the bounds that it had set for itself—the style endured throughout the 16th century and even extended into the 17th. The political subjugation of the country by the forces of Islām may have contributed to the conservatism of the style but did not result in its total elimination, as seems to have been the case in eastern India. Indeed, in the course of its long life, the western Indian school became a national style, painting at other centres in India interpreting and elaborating its forms in their own individual manner. In the province of Orissa, painting on palm leaf and in a manner entirely dependent on the western Indian style has continued up to the present day.

Transition to the Mughal and Rajasthani styles. The belief held earlier by scholars that the new Islāmic rulers of India did not patronize any painting until the rise of the Mughal dynasty in the 16th century is being abandoned in the face of the literary testimony and the discovery or recognition of illustrated manuscripts that were painted at Indian courts. Nor should this be surprising, as the Muslim kings of India had before them the example of other rulers of the Islāmic world who were great patrons of painting in spite of the injunctions of orthodox Islām against the portrayal of living beings. The taste of these Indian rulers, however, did not turn to the western Indian style but to the flourishing traditions of Islāmic painting abroad, notably

Folio from an illustrated manuscript of *Candāyana*, *c.* first half of the 16th century. In the Prince of Wales Museum of Western India, Bombay.

P Chandra

neighbouring Iran. As many painters as architects had in all probability been invited from foreign countries; and illustrated manuscripts, handily transported, must have been easily available. As a result there appears to have developed what can only be called an Indo-Persian style, based essentially on the schools of Iran but affected to a greater or lesser extent by the individual tastes of the Indian rulers and by the local styles. The earliest known examples are paintings dating from the 15th century onward. The most important are the *Khamseh* ("Quintet") of Amīr Khosrow of Delhi (Freer Gallery of Art, Washington, D.C.), a *Bostān* painted in Mandu (National Museum, New Delhi), and, most interesting of all, a manuscript of the *Ne'mat-nāmeh* (India Office Library, London) painted for a sultan of Mālwa in the opening years of the 16th century. Its illustrations are derived from the Turkmen style of Shīrāz but show clear Indian features adapted from the local version of the western Indian style.

P Chandra

Folio from a series illustrating the *Caurapañcāśikā* of Bilhaṇa, *c.* mid-16th century AD. In the Municipal Museum, Ahmadābād, Gujarāt, India.

Though the western Indian style was essentially conservative, it was not unfailingly so. It began to show signs of an inner change most notably in two manuscripts from Mandu, a *Kalpa-sūtra* and a *Kālakācāryakathā* of about 1439, and a *Kalpa-sūtra* painted at Jaunpur in 1465. These works were done in the opulent manner of the 15th century, but for the first time the quality of the line is different, and the uncompromisingly abstract expression begins to make way for a more human and emotional mood. By the opening years of the 16th century, a new and vigorous style had come into being. Although derived from the western Indian style, it is clearly independent, full of the most vital energy, deeply felt, and profoundly moving. The earliest dated example is an *Araṇyaka Parva* of the *Mahābhārata* (1516; The Asiatic Society of Bombay), and among the finest are series illustrating the *Bhāgavata-Purāṇa* and the *Caurapañcāśikā* of Bilhaṇa, scattered in collections all over the world. A technically more refined variant of this style, preferring the pale, cool colours of Persian derivation, a fine line, and meticulous ornamentation, exists contemporaneously and is best illustrated by a manuscript of the ballad *Candāmyana* by Mullā Dāūd (*c.* first half of the 16th century; Prince of Wales Museum of Western India, Bombay). The early 16th century thus appears to have been a period of inventiveness and set the stage for the development of

Fifteenth-century changes in western Indian style

the Mughal and Rājput schools, which thrived from the 16th to the 19th century.

Mughal style: Akbar period (1556–1605). Although the Mughal dynasty came to power in India with the great victory won by Bābar at the Battle of Pānīpat in 1526, the Mughal style was almost exclusively the creation of Akbar. Trained in painting at an early age by a Persian master, Khwāja ʿAbd-uṣ-Ṣamad, who was employed by his father, Humāyūn, Akbar created a large atelier, which he staffed with artists recruited from all parts of India. The atelier, at least in the initial stages, was under the superintendence of Akbar's teacher and another great Persian master, Mīr Sayyid ʿAlī; but the distinctive style that evolved here owed not a little to the highly individual tastes of Akbar himself, who took an interest in the work, inspecting the atelier frequently and rewarding painters whose work was pleasing.

The work of the Mughal atelier in this early formative stage was largely confined to the illustration of books on a wide variety of subjects: histories, romances, poetic works, myths, legends, and fables, of both Indian and Persian origin. The manuscripts were first written by calligraphers, with blank spaces left for the illustrations. These were executed largely by groups of painters, including a colourist, who did most of the actual painting, and specialists in portraiture and in the mixing of colours. Chief of the group was the designer, generally an artist of top quality, who formulated the composition and sketched in the rough outline. A thin wash of white, through which the initial drawing was visible, was then applied and the colours filled in. The colourist's work proceeded slowly, the colour being applied in several thin layers and frequently rubbed down with an agate burnisher, a process that resulted in the glowing, enamel-like finish. The colours used were mostly mineral but sometimes consisted of vegetable dyes; and the brushes, many of them exceedingly fine, were made from squirrel's tail or camel hair.

The technique of Mughal painting

The earliest paintings (*c.* 1560–70) of the school of Akbar are illustrations of *Ṭūṭī-nāmeh* ("Parrot Book; Cleveland Museum of Art) and the stupendous illustrations of the *Dāstān-e Amīr Ḥamzeh* ("Stories of Amīr Ḥamzeh"; Österreichisches Museum für Angewandte Kunst, Vienna), which originally consisted of 1,400 paintings of an unusually large size (approximately 25 inches by 16 inches [65 by 40 centimetres]), of which only about 200 have survived. The *Ṭūṭī-nāmeh* shows the Mughal style in the process of formation: the hand of artists belonging to the various non-Mughal traditions is clearly recognizable, but the style also reveals an intense effort to cope with the demands of a new patron. The transition is achieved in the *Dāstān-e Amīr Ḥamzeh,* in which the uncertainties are overcome in a homogeneous style, quite unlike Persian work in its leaning toward naturalism and filled with swift, vigorous movement and bold colour. The forms are individually modelled, except for the geometrical ornament used as architectural decor; the figures are superbly interrelated in closely unified compositions, in which depth is indicated by a preference for diagonals; and much attention is paid to the expression of emotion. One of the last manifestations of this bold and vigorous early manner is the *Dārābnameh (c.* 1580) in the British Museum.

Immediately following were some very important historical manuscripts, including the *Tārīkh-e Khāndān-e Tīmūrīyeh* ("History of the House of Timūr," *c.* 1580–85; Khuda Baksh Library, Patna) and other works concerned with the affairs of the Tīmūrid dynasty, to which the Mughals belonged. Each of these manuscripts contains several hundred illustrations, the prolific output of the atelier made possible by the division of labour that was in effect. Historical events are recreated with remarkable inventiveness, though the explosive and almost frantic energy of the *Dāstān-e Amīr-Ḥamzeh* has begun to subside. The scale was smaller and the work began to acquire a studied richness. The narrative method employed by these Mughal paintings, like that of traditional literature, is infinitely discursive; and the painter did not hesitate to provide a fairly detailed picture of contemporary life—both of the people and of the court—and of the rich fauna and flora of India. Like Indian artists of all periods, the Mughal painter showed a remarkable empathy for animals, for through them flows the same life that flows through human beings. This sense of kinship allowed him to achieve unqualified success in the illustration of animal fables such as the *Anwār-e Suhaylī* ("Lights of Canopus"), of which several copies were painted, the earliest dated 1570 (School of Oriental and African Studies, London). It was in the illustrations to Persian translations of the Hindu epics, the *Mahābhārata* and the *Rāmāyaṇa,* that the Mughal painter revealed to the full the richness of his imagination and his unending resourcefulness. With little precedent to rely on, he was nevertheless seldom dismayed by the subject and created a whole series of convincing compositions. Because most of the painters of the atelier were Hindus, the subjects must have been close to their hearts; and, given the opportunity by a tolerant and sympathetic patron, they rose to great heights. It is no wonder, therefore, that the *Razm-nāmeh* (City Palace Museum, Jaipur), as the *Mahābhārata* is known in Persian, is one of the outstanding masterpieces of the age.

Illustrations for Persian translations of Hindu epics

In addition to large books containing numerous illustrations, which were the products of the combined efforts of many artists, the imperial atelier also cultivated a more intimate manner that specialized in the illustration of books, generally poetic works, with a smaller number of illustrations. The paintings were done by a single master artist who, working alone, had ample scope to display his virtuosity. In style the works tend to be finely detailed and exquisitely coloured. A *Dīvān* ("Anthology") of Anwarī (Fogg Art Museum, Cambridge, Massachusetts), dated 1589, is a relatively early example of this manner. The paintings are very small, none larger than five inches by $2^1/_2$ inches (12 by 6 centimetres) and most delicately executed. Very similar in size and quality are the miniatures illustrating the *Dīvān* of Ḥāfeẓ (Reza Library, Rāmpur). On a larger scale but in the same mood are the manuscripts that represent the most delicate and refined works of the reign of Akbar: the *Bahāristān* of Jāmī (1595; Bodleian Library, Oxford), a *Khamseh* of Neẓāmī (1593; British Museum, London), a *Khamseh* of Amīr Khosrow (1598; Walters Art Gallery, Baltimore and Metropolitan Museum of Art, New York), and an *Anwār-e Suhaylī* (1595–96; Bharat Kala Bhavan, Vārānasī).

Also prepared in the late 1590s were magnificent copies

P. Chandra

Painter at work, detail from a folio of the Muraqqah-e Gulshan, Mughal style, early 17th century AD. In the Staatliche Museen Preussischer Kulturbesitz, Berlin.

Nobleman seated on a terrace, Mughal style painting, mid-18th century. In the National Museum of India, New Delhi.

P Chandra

of the *Akbar-nāmeh* ("History of Akbar"; Victoria and Albert Museum, London) and the *Kitāb-e Changīz-nāmeh* ("History of Genghis Khan"; Gulistan Library, Tehrān). These copiously illustrated volumes were produced by artists working jointly, but the quality of refinement is similar to that of the poetic manuscripts.

Of the large number of painters who worked in the imperial atelier, the most outstanding were Dasvant and Basāvan. The former played the leading part in the illustration of the *Razm-nāmeh*. Basāvan, who is preferred by some to Dasvant, painted in a very distinctive style, which delighted in the tactile and the plastic, and with an unerring grasp of psychological relationships.

Mughal style: Jahāngīr period (1605–27). The emperor Jahāngīr, even as a prince, showed a keen interest in painting and maintained an atelier of his own. His tastes, however, were not the same as those of his father, and this is reflected in the painting, which underwent a significant change. The tradition of illustrating books began to die out, though a few manuscripts, in continuation of the old style, were produced. For Jahāngīr much preferred portraiture; and this tradition, also initiated in the reign of his father, was greatly developed. Among the most elaborate works of his reign are the great court scenes, several of which have survived, showing Jahāngīr surrounded by his numerous courtiers. These are essentially large-scale exercises in portraiture, the artist taking great pains to reproduce the likeness of every figure.

Portraiture

The compositions of these paintings have lost entirely the bustle and movement so evident in the works of Akbar's reign. The figures are more formally ordered, their comportment in keeping with the strict rules of etiquette enforced in the Mughal court. The colours are subdued and harmonious, the bright glowing palette of the Akbarī artist having been quickly abandoned. The brushwork is exceedingly fine. Technical virtuosity, however, is not all that was attained, for beneath the surface of the great portraits of the reign there is a deep and often spiritual understanding of the character of the person and the drama of human life.

Many of the paintings produced at the imperial atelier are preserved in the albums assembled for Jahāngīr and his son Shāh Jahān. The Muraqqah-e Gulshan is the most spectacular. (Most surviving folios from this album are in the Gulistan Library in Tehrān and the Staatliche Museen Preussischer Kulturbesitz, Berlin; a section is temporarily housed in Tübingen.) There are assembled masterpieces from Iran, curiosities from Europe, works produced in the reign of Akbar, and many of the finest paintings of Jahāngīr's master painters, all surrounded by the most magnificent borders decorated with a wide variety of floral and geometrical designs. The album gives a fairly complete idea of Jahāngīr as a patron, collector, and connoisseur of the arts, revealing a person with a wide range of taste and a curious, enquiring mind.

Jahāngīr esteemed the art of painting and honoured his painters. His favourite was Abū al-Hasan, who was designated Nādir-uz-Zamān ("Wonder of the Age"). Several pictures by the master are known, among them a perceptive study of Jahāngīr looking at a portrait of his father. Also much admired was Ustād Manṣūr, designated Nādir-ul-ʿAṣr ("Wonder of the Time"), whose studies of birds and animals are unparalleled. Bishandās was singled out by the emperor as unique in the art of portraiture. Manohar, the son of Basāvan, Govardhan, and Daulat are other important painters of this reign.

Mughal style: Shāh Jahān period (1628–58). Under Shāh Jahān, attention seems to have shifted to architecture, but painting in the tradition of Jahāngīr continued. The style, however, becomes noticeably rigid. The portraits resemble hieratic effigies, lacking the breath of life so evident in the work of Jahāngīr's time. The colouring is jewel-like in its brilliance, and the outward splendour quite dazzling. The best work is found in the *Shāhjahānnāmeh* ("History of Shāh Jahān") of the Windsor Castle Library and in several albums assembled for the emperor. Govardhan and Bichitra, who had begun their careers in the reign of Jahāngīr, were among the outstanding painters; several

P. Chandra

The musical mode *vasanta*, Deccani school painting, Bijāpur, late 16th century. In the National Museum of India, New Delhi.

works by them are quite above the general level produced in this reign.

Mughal style: Aurangzeb and the later Mughals (1659–1806). From the reign of Aurangzeb (1659–1707), a few pictures have survived that essentially continue the cold style of Shāh Jahān; but the rest of the work is nondescript, consisting chiefly of an array of lifeless portraits, most of them the output of workshops other than the imperial atelier. Genre scenes, showing gatherings of ascetics and holy men, lovers in a garden or on a terrace, musical parties, carousals, and the like, which had grown in number from the reign of Shāh Jahān, became quite abundant. They sometimes show touches of genuine quality, particularly in the reign of Muḥammad Shāh (1719–48), who was passionately devoted to the arts. This brief revival, however, was momentary, and Mughal painting essentially came to an end during the reign of Shāh ʿĀlam II (1759–1806). The artists of this disintegrated court were chiefly occupied in reveries of the past, the best work, for whatever it is worth, being confined to copies of old masterpieces still in the imperial library. This great library was dispersed and destroyed during the uprising of 1857 against the British.

Genre scenes

Company school. Rising British power, which assumed political supremacy in the 19th century, resulted in a radical change of taste brought about by the Westernization of important segments of the population. Heavily influenced by Western ideas, a style emerged that represented the adjustment of traditional artists to new fashions and demands. Rooted at Delhi and the erstwhile provincial Mughal capitals of Murshidābād, Lucknow, and Patna, it ultimately spread all over India. Most of the works produced were singularly impoverished, but occasionally there were some fine studies of natural life.

Deccani style. In mood and manner, Deccani painting, which flourished over much of the Deccan Plateau from at least the last quarter of the 16th century, is reminiscent of the contemporary Mughal school. Again, a homogeneous style evolved from a combination of foreign (Persian and Turkish) and Indian elements, but with a distinct local flavour. Of the early schools, the style patronized by the sultans of Bijāpur—notably the tolerant and art-loving Ibrāhīm ʿĀdil Shāh II of Bijāpur, famous for his love of music—is particularly distinguished. Some splendid portraits of him, more lyrical and poetic in concept than contemporary Mughal portraits, are to be found. A wonderful series depicting symbolically the musical modes (*rāgamālā*) also survives. Of illustrated manuscripts, the most important are the *Nujūm-ul-ʿulūm* ("The Stars of the Sciences," 1590; Chester Beatty Library, Dublin) and the *Tārīf-e Ḥuseyn-Shāhī* (Bharata Itihasa Samshodhaka Mandala, Pune), painted around 1565 in the neighbouring state of Ahmadnagar. The sultanate of Golconda also produced work of high quality—for example, a manuscript of the *Dīvān* of Muḥammad Qulī Quṭb Shāh in the Salar Jang Library, Hyderābād, and a series of distinguished portraits up to the end of the 17th century (dispersed in various collections). The state of Hyderābād, founded in the early 18th century and headed by a grandee of the Mughal Empire, was a great centre of painting. The work that was produced there reflects both Golconda traditions and increasing Mughal and Rajasthani influences.

Rajasthani style. This style appears to have come into being in the 16th century, about the same time the Mughal school was evolving under the patronage of Akbar; but, rather than a sharp break from the indigenous traditions, it represented a direct and natural evolution. Throughout the early phase, almost up to the end of the 17th century, it retained its essentially hieratic and abstract character, as opposed to the naturalistic tendencies cultivated by the Mughal atelier. The subject matter of this style is essentially Hindu, devoted mainly to the illustration of myths and legends, the epics, and above all the life of Krishna; particularly favoured were depictions of his early life as the cowherd of Vraja, and the mystical love of Vraja's maidens for him, as celebrated in the *Bhāgavata-Purāṇa*, the *Gītagovinda* of Jayadeva, and the Braj Bhasa verses written by Sūrdās and other poets. The style of the painting, no less than the literature, is a product of the new religious movements, all of which stressed personal devotion to Krishna as the way to salvation. Related popular themes were pictorial representations of the musical modes (*rāgamālā*) and illustrations of poetical works such as the *Rasikapriyā* of Keśavadāsa, which dealt with the sentiment of love, analyzing its varieties and endlessly classifying the types of lovers and beloveds and their emotions. Portraits, seldom found in the early phase, became increasingly common in the 18th century—as did court scenes, scenes of sporting and hunting events, and other scenes concerned with the courtly life of the great chiefs and feudal lords of Rājasthān.

The Hindu basis of Rajasthani style

The Rajasthani style developed various distinct schools, most of them centring in the several states of Rājasthān, namely Mewār, Būndi, Kotah, Mārwār, Bīkaner, Kishangarh, and Jaipur (Amber). It also had centres outside the geographical limits of present-day Rājasthān, notably Gujarāt, Mālwa, and Bundel Khand. The study of Rajasthani painting is still in its infancy, for most of the material has been available for study only since the mid-1940s.

The Mughal and Rajasthani styles were always in contact with each other, but in general the Rajasthani schools were not essentially affected by the work produced at the Mughal court during the greater part of the 17th century. This became less true in the 18th century, when the sharp distinction between the two became progressively obscured, though each retained its distinctive features right up to the end.

Rajasthani style: Mewār. The Mewār school is among the most important. The earliest dated examples are represented by a *rāgamālā* series painted at Chawand in 1605 (Gopi Krishna Kanoria Collection, Patna). These simple paintings, filled with bright colour, are only a step removed from the pre-Rajasthani phase. The style became more elaborate in the first quarter of the 17th century when another *rāgamālā*, painted at Udaipur in 1628 (formerly in the Khajanchi Collection, Bīkaner; now dispersed in various collections), showed some superficial acquaintance with the Mughal manner. This phase, lasting until around 1660, was one of the most important for the development of painting all over Rājasthān. Ambitious and extensive illustrations of the *Bhāgavata*, the *Rāmāyaṇa*, the poems of Sūrdās, and the *Gītagovinda* were completed, all full of strength and vitality. The name of Sāhabadī is intimately connected with this phase; another well-known painter is Manohar. The intensity and richness associated with their

P Chandra

Lion hunt, Rajasthani style painting, Kotah, late 18th century. In a private collection.

atelier began to fade toward the close of the 17th century, and a wave of Mughal influence began to affect the school in the opening years of the 18th century. Portraits, court scenes, and events in the everyday world of the ruling classes are increasingly found. Although the emotional fervour of the 17th century was never again attained, this work is often of considerable charm. The 19th century continued to create work in the tradition of the 18th, one of the most important centres being Nāthdwāra (Rājasthān), the seat of the Vallabha sect. Large numbers of pictures, produced here for the pilgrim trade, were spread over all parts of Rājasthān, northern India, Gujarāt, and the Deccan.

Earliest examples of Būndi school

Rajasthani style: Būndi and Kotah. A school as important as that of Mewār developed at Būndi and later at Kotah, which was formed by a partition of the parent state and ruled by a junior branch of the Būndi family. The earliest examples are represented by a *rāgamālā* series of extraordinarily rich quality, probably dating from the end of the 16th century. From the very beginning the Būndi style seemed to have found Mughal painting an inspiring source, but its workmanship was as distinctively Rajasthani as the work of Mewār. The artists of this school always displayed a pronounced preference for vivid movement, which is unique in all of Rājasthān. Toward the second half of the 17th century, work at Būndi came unmistakably under the influence of Mewār; many miniatures, including several series illustrating the *Rasikapriyā,* indicate that this was a period of prolific activity. The sister state of Kotah also appears to have become an important centre of painting at this time, developing a great fondness for hunting and sport scenes, all filled with great vigour and surging strength. This kind of work continued well into the 19th century, and if the workmanship is not always of the highest quality, the style maintained its integrity against the rapidly increasing Western influence right up to the end.

Rajasthani style: Mālwa. It has been suggested but not definitely determined that the school itself does not belong to Mālwa but to some other area, probably Bundelkhand. In contrast to the Būndi school, miniatures generally thought to have been painted in Mālwa are quite archaistic, with mannerisms inherited from the 16th century retained until the end of the 17th. The earliest work is an illustrated version of the *Rasikapriyā* (1634), followed by a series illustrating a Sanskrit poem called the *Amaru Śataka* (1652). There are also illustrations of the musical modes (*rāgamālā*), the *Bhāgavata-Purāṇa,* and other Hindu devotional and literary works. The compositions of all of these pictures is uncompromisingly flat, the space divided into registers and panels, each filled with a patch of colour and occupied by figures that convey the action. This conservative style disappeared after the close of the 17th century. The course of Mālwa painting in the 18th century and later is not known.

P. Chandra

A priest at worship, Rajasthani style painting, Nāthdwāra, mid-19th century. In a private collection.

A Rājput king listening to music, Rajasthani style painting, Mārwār, mid-18th century. In a private collection.

P. Chandra

Rajasthani style: Mārwār. A *rāgamālā* series dated 1623 reveals that painting in this state shared features common to other schools of Rājasthān. Miniatures of the second half of the 17th century are distinguished by some splendid portraits that owe much to the Mughal school. A very large amount of work was done in the 19th century, all of which is highly stylized but strong in colour and often of great charm.

Rajasthani style: Bīkaner. Of all Rajasthani schools, the Bīkaner style, from its very inception in the mid-17th century, shows the greatest indebtedness to the Mughal style. This is due to the presence in the Bīkaner atelier of artists who had previously worked in the Mughal manner at Delhi. They and their descendants continued to paint in a style that could only be classed as a provincial Mughal manner had it not been for the quick absorption of influences from the Rajasthani environment and a sympathy for the religious and literary themes favoured by the royal Hindu patrons. Delicacy of line and colour are consistent characteristics of Bīkaner painting even when, toward the end of the 18th century, it assumed stylistic features associated with the more orthodox Rajasthani schools.

Characteristics of the Bīkaner style

Rajasthani style: Kishangarh. The Kishangarh school, which came into being toward the mid-18th century, was also indebted to the contemporary Mughal style but combined a rich and refined technique with deeply moving religious fervour. Its inspiring patron in the formative phase was Sāvant Singh, more of a devotee and a poet than a king. The style established by him, characterized by pronounced mystical leanings and a distinctive facial type, continued to the middle of the 19th century, though at a clearly lower level of achievement.

Rajasthani style: Jaipur (Amber). The rulers of the state

Month of summer, Rajasthani-style painting, Bīkaner, early 18th century. In a private collection.
P. Chandra

were closely allied to the Mughal dynasty, but paintings of the late 16th and early 17th centuries possessed all of the elements of the Rajasthani style. Little is known about the school until the opening years of the 18th century, when stiff, formal examples appear in the reign of Savāī Jai Singh. The finest works, dating from the reign of Pratāp Singh, are sumptuous in effect and include some splendid portraits and some large paintings of the sports of Krishna. Although the entire 19th century was extremely productive, the work was rather undistinguished and increasingly affected by Western influences. Of the Rajasthani styles of this period, the Jaipur school was the most popular, examples having been found all over northern India.

Pahari style. Closely allied to the Rajasthani schools both in subject matter and technique is the Pahari style, so-named because of its prevalence in the erstwhile hill states of the Himalayas, stretching roughly from Jammu to Gaṛhwāl. It can be divided into two main schools, the Basohlī and the Kāngra, but it must be understood that these schools were not confined to the centres after which they are named but extended all over the area. Unlike Rājasthān, the area covered by the Pahari style is small, and the probability of artists travelling from one area to another in search of livelihood was much greater. Thus, attempts to distinguish regional schools are fraught with controversy, and it has been suggested that a classification based upon ateliers and families is likely to be more tenable than those presently current among scholars. Because the Basohlī and the Kāngra schools show considerable divergences, scholars have postulated the existence of a transitional phase, named the pre-Kāngra style.

Pahari style: Basohlī school. The origins of this remarkable style are not yet understood, but it is clear that the style was flourishing toward the close of the 17th century. The earliest dated paintings are illustrations to the *Rasamañjarī* of Bhānudatta (a Sanskrit work on poetics), executed for a ruler of Basohlī (1690; Boston Museum of Fine Arts). Bold colour, vigorous drawing, and primitive intensity of feeling are outstanding qualities in these paintings, quite surpassing the work of the plains. In addition to other Hindu works such as the *Gītagovinda* and the *Bhāgavata-Purāṇa,* a fairly large number of idealized portraits have also been discovered.

Pahari style: Kāngra school. The Basohlī style began to fade by the mid-18th century, being gradually replaced by the Kāngra style, named after the state of Kāngra but, like the Basohlī style, of much wider prevalence. A curvilinear line, easy flowing rhythms, calmer colours, and a mood of sweet lyricism easily distinguish the work from that of the Basohlī style. The reasons for this change are to be sought in strong influences from the plains, notably the Mughal styles of Delhi and Lucknow. These influences account for the more refined technique; but whatever was borrowed was transmuted and given a fresh and tender aspect. Among the greatest works are large series illustrating the *Bhāgavata-Purāṇa* (National Museum, New Delhi), the *Gītagovinda,* and the *Satsaī* of Bihārī (both in the collection of the maharaja of Tehrī-Garhwāl), all painted in 1775–80. The corpus of work produced is very large and, although it seldom fails to please, works of high achievement are rare. The school flourished from about 1770 to almost the end of the 19th century, but the finest work was produced largely around 1775–1820.

Modern period. Toward the late 19th century traditional Indian painting was rapidly dying out, being replaced by feeble works in a variety of idioms, all strongly influenced by the West. A reaction set in during the early 20th century, symbolized by what is called the Bengal school. The glories of Indian art were rediscovered, and the school consciously tried to produce what it considered a truly Indian art inspired by the creations of the past. Its leading artist was Rabindranath Tagore and its theoretician was E.B. Havell, the principal of the Calcutta School of Art. Nostalgic in mood, the work was mainly sentimental though often of considerable charm. The Bengal school did a great deal to reshape contemporary taste and to make Indian artists aware of their own heritage. Amrita Sher-Gil, who was inspired by the Postimpressionists, made Indian painters aware of new directions. Mid-20th-century Indian painting is very much a part of the international scene, the artists painting in a variety of idioms, often attempting to come to terms with their heritage and with the emergence of India as a modern culture.

The Bengal school

Decorative arts. Fragmentary ivory furniture (*c.* 1st century AD) excavated at Begrām is one of the few indications of the existence in ancient India of a secular art concerned with the production of luxurious and richly decorated objects meant for daily use. Objects that can be clearly designated as works of decorative art become much more extensive for the later periods, during which Islāmic traditions were having a profound effect on Indian artistic traditions. The reign of the Mughal emperors, in particular, produced works of the most elaborate and exquisite craftsmanship; the decorative tradition is clearly preserved in architectural ornament, though surviving decorative objects themselves, particularly before the 17th century, are far fewer than might be expected. Economic conditions, including competition with machine-made goods imported from English factories, and a change in taste from increasing European influence had disastrous consequences for traditional craftsmanship, especially in the late 19th and 20th centuries.

Pre-Islāmic period. Of the very few objects surviving from the pre-Islāmic period, the most important are fragments of ivory caskets, chairs, and footstools found at Begrām, in eastern Afghanistan, but obviously of Indian origin and strongly reminiscent of the school of Mathurā in the 1st century AD. The work is profusely decorated with carved panels and confirms the wide reputation for superb ivories that India had in ancient times. Nothing as spectacular has come down from the succeeding periods, but stray examples such as the so-called Charlemagne chessman (*c.* 8th century; Cabinet des Medailles, Paris) and two magnificent throne legs, of Orissan workmanship, carved in the shape of griffins with elephant heads (13th century; Freer Gallery of Art, Washington D.C., and Philadelphia Museum of Art), indicate that ivory craftsmanship was always vital. Ancient traditions, relatively unaffected by Islāmic influence, continued in southern India up to modern times. An exquisitely carved box from Vijayanagar (16th century; Prince of Wales Museum of Western India, Bombay) is representative; many other exquisite objects of this period and later are among the treasures of South Indian temples.

Work in ivory

Ritual vessel from Kollur, Mysore, India, copper, *c.* 14th century. In the Prince of Wales Museum of Western India, Bombay.
P. Chandra

There is even less evidence of what the decorative work in metal was like. The practice of re-using the metal by melting unserviceable items may account for the paucity of objects, for there is little doubt that the craft was always flourishing. A hoard found at Kolhāpur, consisting of plates, various kinds of vessels, lamps and *objets d'art,* including a superb bronze elephant with riders, constitutes the most important surviving group of metal objects and is datable to about the 2nd century AD. Some fine examples of ritual utensils, notably elaborate incense burners, of the 8th–9th century have been excavated at Nālandā; and a large number of 14th-century ceremonial vessels of complex design and excellent workmanship, and apparently belonging to the local temple, were discovered at Kollur, in Mysore state.

Gold played an extremely important role in the manufacture of jewelry, but once again the finds are hardly commensurate with tradition. Small amounts of gold jewelry have been excavated at Mohenjo-daro and Harappā (3rd millennium BC); and, in the historical period, a very important group, of delicate workmanship, has been excavated at Taxila (*c.* 2nd century AD).

From earliest times, India has been famous for the variety and magnificence of its textiles. In this case, however, the Indian climate has been particularly destructive; virtually nothing has survived the heat and moisture. Besides the testimony of literature and the evidence of figural sculpture, only a few fragments of printed textiles are preserved—at Fusṭāṭ in Egypt, where they had been exported. These date approximately to the 14th century.

Islāmic period. Traditions of craftsmanship established during the Islāmic period came to full flower during the reign of the Mughal dynasty. Surviving works of decorative art are more abundant, though once again there are hardly as many examples as might be expected, particularly from the 16th and 17th centuries. According to literary testimony and the few available examples, the finest objects were undoubtedly made in the imperial workshops set up in large number at the capital and in the great cities of the empire, where they were nourished by local traditions. Well-organized, these shops specialized in particular items, such as textiles, carpets, jewelry, ornamental arms and armour, metalware, and jade. Textile manufacture must have been enormous, considering the demands of court and social etiquette and ritual. Contributing to the popularity of tapestries, curtains, draperies, canopies, and carpets in contemporary architecture were the nomadic tenting traditions of the Mughal rulers.

Popularity of textiles

P. Chandra

Detail of a Kashmir shawl, wool embroidery, 19th century. In the Prince of Wales Museum of Western India, Bombay.

The variety of techniques employed in the manufacture of textiles was infinite, ranging from printed and painted patterns to the exquisite embroidery decoration of woolen shawls and the costly figured brocading of silk. An important contribution to carpet weaving was the landscape carpet that reproduced pictorial themes inspired by miniature painting. Much of the surviving textile work dates from the 18th century or later, though the 16th and 17th centuries produced works of the most outstanding quality.

In response to growing European trade, a considerable amount of furniture (chairs, cabinets, chests of drawers, and the like) was produced, mostly wood inlaid with ivory. Many of these pieces have been preserved in the kinder European climate. Although the furniture made for export gives some idea of the craft in India, it must be emphasized that only the ornamental and figural work was Indian, while the form was European. Also in a hybrid Indo-European style were the Christian objects produced by a local school of ivory carvers at Goa.

Metal objects of sumptuous quality were also made, a unique example of which is a splendid, elaborately chiselled 16th-century cup in the Prince of Wales Museum of Western India in Bombay. This tradition was continued in the 17th and particularly the 18th century, when vessels made of a variety of metals and adorned with engraved, chiselled, inlaid, and enamelled designs were very popular. Arms and armour, in particular, were decorated with the skill of a jeweler. Particularly striking are the carved hilts, often done in animal shapes.

Jade or jadeite was much fancied by the rich and was used together with crystal to make precious vessels as well as sword and dagger hilts. A rather large number of 18th- and 19th-century objects have survived, but they are often of nondescript quality. The greatest period for jade carving seems to have been the 17th century; a few outstanding examples associated with the emperors Jahāngīr and Shāh Jahān are of singular delicacy and perfection. The practice of inlaying jade, and also stone, with precious or semiprecious stones became more popular with the reign of Shāh Jahān and increasingly characteristic of Indian jade craftsmanship from that time on.

Architectural decoration provides a clear idea of the range of ornamental patterns used by the Mughal artist. They consisted mainly of arabesques (intricate interlaced patterns made up of flower, foliage, fruit, and sometimes animal and figural outlines) and infinitely varied geometric patterns—motifs shared with the rest of the Muslim world—together with floral scrolls and other designs adapted from Indian traditions. As a whole, the Mughal decorative style tends to endow ornamental patterns with a distinctive plasticity not seen in the more truly two-dimensional Iranian and Arab work. From the 17th century, a type of floral spray became the most favoured motif and was found on almost every decorated object. The motif, symmetrical but relatively naturalistic at the beginning, became progressively stiff and stylized, but never lost its importance in the ornamental vocabulary.

VISUAL ARTS OF SRI LANKA (CEYLON)

The art of Sri Lanka is closely allied to that of India but presents several distinctive features that make a separate treatment convenient. There is, first, the considerable transformation of Indian influences, resulting in an idiom of great power and individuality. Sri Lanka also often served as a geographical pocket in which styles that had disappeared in India were preserved, which accounts for the anachronistic features of some phases of Sinhalese art. It also appears that although predominant influences were from neighbouring southern India, this was not exclusively

Distinction between art of Sri Lanka and that of India

Abhayagiri *dagāba* in the Jetavana at Anurādhapura, Sri Lanka, *c.* 4th century.
ZEFA

so, and styles flourishing in western and northern India, too, contributed to the formulation of Sinhalese art. The difficulties in the study of the art are considerable: the long, unbroken Buddhist traditions and the piety of the rulers and the people have led to the successive renovation of monuments; and in the absence of firmly dated monuments, one of the few relatively reliable tools of study is comparison with Indian art, an approach full of pitfalls and shortcomings.

Architecture. The most impressive monuments are the great *stūpa*s, some of gigantic size and considerable antiquity but often reconstructed in the course of the centuries. They generally have a triple circular base, and as in early Indian *stūpa*s, a hemispherical dome with a miniature railing on top, and a multiple parasol that tends to solidify into a conical structure in the course of time. The material is brick, sometimes covered with plaster and white paint. An important feature are the platforms (*vāhalakaḍa*s) at the cardinal points, often adorned with sculpture. There are many *stūpa*s at the ancient capital of Anurādhapura, at Polonnaruva, and at other sites; of these the Jetavana at Anurādhapura is the largest, though now largely ruined.

ZEFA

The *vaṭadāgē* at Polonnaruva, Sri Lanka, 12th century.

Small *stūpa*s were often placed in a circular building with a domical metal and timber roof supported by concentric rows of stone pillars. This type of building, known in ancient India as the *caityagṛha,* was very popular in Sri Lanka, though it had disappeared at a fairly early period in the country of its origin. A famous example is the *vaṭadāgē* at Polonnaruva, a structure of great elegance. The dome itself, being of perishable material, has not survived. The *geḍigē,* or large rectangular hall with a corbelled brick vault, housing a Buddha image, is first found in Sri Lanka from the 8th century AD; the most impressive example is the Laṅkātilaka at Polonnaruva built by Parākramabāhu I in the 12th century.

Literature testifies to the existence of elaborate royal and priestly residences of wood, which have largely disappeared. The Lohapāsāda at Anurādhapura, traditionally ascribed to Duṭṭhagāmaṇi (101–77), was originally a nine-story building, now destroyed except for the large number of stone pillars that supported the upper floors. Sigiriya, a 6th-century fortress city with extensive remains, is another notable example of secular architecture.

Sculpture. The earliest sculpture, perhaps, is from the platforms, or *vāhalakaḍa*s, of the Kaṇṭaka Cetiya, at Mihintalē, and reveals an archaistic style indebted to 1st-century-BC Indian sculpture of Sānchi and Amarāvatī regions. A certain simplicity and restraint characteristic of most Sinhalese work is present even at this early stage. The first Buddha images show a pronounced relationship to examples from Andhradeśa of the 2nd–3rd century AD but often possess considerable vigour, revealing the contribution of the local sculptor. Several fine images are known, one of the best of which is at Ruanveli, Anurādhapura, now very badly restored.

Dated monuments are absent from the 5th to the 12th centuries, but an approximate idea of stylistic development can be obtained by a comparative study of Indian examples. An outstanding image, rather hideously repaired in recent years, is a great seated Buddha in Anurādhapura, the smooth and abstract modelling of which recalls the school of Sārnāth of the 5th–6th century. At Isurumuni, near Anurādhapura, are some marvellous reliefs carved on rocks. One of these depicts elephants at play, and another, a seated man with the head of a horse carved in the background. These fine sculptures recall the South Indian style of the 7th century. A radiant amorous couple carved

Nāga stele from Anurādhapura, Sri Lanka, 10th century.
George Holton—Photo Researchers

in relief on a stone slab, also at Isurumuni, represents Sinhalese sculpture at its most joyous.

Sculptured staircases

Of about the same period or a little later, are exquisitely sculptured staircases decorated with moonstones, and stelae, or commemorative pillars, carved with a guardian *nāga,* a spirit with combined superhuman and serpent qualities. The latter are among the finest examples of Sinhalese sculpture, the full and weighty modelling relieved by the skillful movement of clearly chiselled ornament. The Ratnapāsāda at Anurādhapura and the eastern staircase of the *vaṭadāgē* at Polonnaruva possess particularly superb specimens. Moonstones—decorated with bands of floral motifs, geese, and a row of animals consisting of a lion, bull, elephant, and horse—placed at the bottom of the staircase, testify to the great taste and elegance that mark Sinhalese decorative carving. At Anurādhapura and related sites a certain freedom characterizes the work, while the slightly later examples at Polonnaruva are stiffer but technically brilliant.

P. Chandra

Painted figure of an *apsaras* from Sigiriya, Sri Lanka, 6th century.

A colossal Buddha, 42 feet (13 metres) high, at Avukana, testifies to the increasing hardness of the Sinhalese style, which, even so, never ceases to be moving. Large images of the Buddha at the Gal Vihāra and a figure supposedly representing Parākramabāhu at Potgal Vihāra, both at Polonnaruva, are of the 12th century. They are figures of great majesty and surpass contemporary work in southern India. After the 13th century, Sinhalese sculpture began to decline, though work of some decorative value was produced up to the 19th century.

Painting. The rock at Sigiriya is adorned with a series of exquisitely painted *apsaras*es (nymphs) showering flowers, their torsos emerging from clouds. The paintings are dated to the 6th century AD; in their plastic resiliency they are reminiscent of contemporary work in India. The next important group of wall paintings come from Tivaṁkapatimā-ghara at Polonnaruva. Although dated to the 12th or 13th century, the figures continue to be modelled, relatively unaffected by the linear distortions of the western Indian style that was flourishing in India. Eighteenth-century paintings, with their flat figures arranged in horizontal rows, reflect contemporary styles of southern India.

(P.Ch.)

BIBLIOGRAPHY

Literature: (Sanskrit, Pāli, and Prākrit): The old literature of Southeast Asia has been more extensively described than any of the modern literatures. Though antiquated, MORIZ WINTERNITZ, *Geschichte der indischen litteratur,* 3 vol. (1908–22; Eng. trans., *A History of Indian Literature,* 2nd ed. 1959–67), is still very useful. So is ARTHUR B. KEITH, *A History of Sanskrit Literature* (1928, reprinted 1956), which, however, does not include the Sanskrit theatre. For the old literature of the Veda, ARTHUR A. MACDONNELL, *A History of Sanskrit Literature* (1900), remains very helpful as a survey. A full inventory of the literature is given by S.N. DAS GUPTA and S.K. DE, *A History of Sanskrit Literature, Classical Period,* 2nd ed. (1962). The epochal material is best treated by EDWARD W. HOPKINS, *The Great Epic of India* (1901). On the Sanskrit play specifically, the study by the French scholar SYLVAIN LEVI, *Le Théâtre indien* (1890), remains an important contribution. The best introduction to the narrative literature is to be found in CHARLES H. TAWNEY (trans.), *The Ocean of Story,* ed. by NORMAN M. PENZER, 10 vol. (1923–28, reprinted 1968). GEORGE L. HART, *The Relation Between Tamil and Classical Sanskrit Literature* (1976), is an argument that the two literatures stem from a common source.

Modern Indian literatures: The literatures in the modern Indian languages are bibliographically underrepresented in English. For some of them the best sources are to be found written in those languages themselves. The following bibliography must confine itself to works written in more accessible languages. (*Hindi*): A good though incomplete inventory of the older literature in Hindi and Hindustani is found in JOSEPH GARCIN DE TASSY, *Histoire de la littérature hindouie et hindoustanie,* 2nd ed., 3 vol. (1870–71, reprinted 1968). A useful guide up to its date is EDWIN GREAVES, *A Sketch of Hindi Literature* (1918). A survey of the more modern literature in Hindi is R.A. DWIVELDI, *A Critical Survey of Hindi Literature* (1966). (*Assamese*): Among the few works available is B.K. BARUA, *Assamesè Literature* (1941) and *A History of Assamese Literature* (1965). (*Bengali*): The best extensive introduction to the literature in Bengali is that of SUKUMAR SEN, *History of Bengali Literature* (1960). Another useful source is J.C. GHOSH, *Bengali Literature* (1948). (*Marathi*): Many works on Marathi literature are written in Marathi itself. A good guide is G.C. BHATE, *History of Modern Marathi Literature, 1800–1938* (1939). (*Gujarati*): Little is written in English on Gujarati literature. To be recommended is K.M. MUNSHI, *Gujarāt and Its Literature from Early Times to 1852,* 3rd ed. (1967). A useful older study is K.M. JHAVERI, *Milestones in Gujarati Literature* (1914). (*Punjabi*): A good survey of Punjabi is given by MOHAN SINGH, *An Introduction to Panjabi Literature* (1951). For information about literature in the smaller Indo-Aryan languages, see the symposium of the All-India Writers' Conference, *Writers in Free India* (1950). (*Muslim contributions*): The best short introduction to the cultural and intellectual life of the Moslems of the subcontinent is AZIZ AHMAD, *An Intellectual History of Islam in India* (1969). A comprehensive survey of the literature in Urdu produced in South India, especially in the medieval kingdoms of Bijāpur and Golconda and in the later state Hyderābād is the Urdu work *Dekan men Urdu* by NASIRRUDDIN HASHMI (1963). A study of the convention of classical Urdu poetry in the light of the writings of three Urdu poets of 18th-century Delhi is RALPH RUSSELL and KHURSHIDUL ISLAM, *Three Mughal Poets: Mir, Sauda, Mir Hasan* (1968). The latest and most comprehensive survey of literature produced in Persian in various countries,

including India and Pakistan, is JAN RYPKA, *History of Iranian Literature* (1968). A useful introduction is MUHAMMAD SADIQ, *A History of Urdu Literature* (1964). (*Tamil*): KAMIL V. ZVELEBIL, *Tamil Literature* (1974), is an introduction to and critical study of the literature. Much useful information is found in XAVIER S. THANI NAYAGAM, *A Reference Guide to Tamil Studies* (1966). Fuller treatment is given in C. and H. JESUDASAN, *A History of Tamil Literature* (1961). Especially recommended are also T.P. MEENAKSHISUNDARAM, *A History of Tamil Literature* (1965); and J.M. SOMASUNDARAM PILLAI, *A History of Tamil Literature with Texts and Translations from the Earliest Times to 600 A.D.* (1968). A good account of Telugu literature is P. CHENCHAYYA and R.M. BHUJANGA RAO BHADUR, *A History of Telugu Literature* (1928). More recent is P.T. RAJU, *Telugu Literature* (1944); and GIDUGU VENKAJA SITAPATI, *History of Telugu Literature* (1968). Very little has been written on Kannada literature. Mention can be made of the older EDWARD P. RICE, *A History of Kanarese Literature*, 2nd ed. rev. and enl. (1921); and of H. THIPPERUDA SWAMY, *The Vīraśaiva Saints* (1968). The literature in Malayalam is sketched in K.M. GEORGE, *A Survey of Malayalam Literature* (1968). Further mention can be made of K.K. NAIR, *A History of Malayalam Literature* (1971); and of P.K. PARAMESWARAN NAIR, *History of Malayalam Literature* (Eng. trans. 1967). Finally, among the very few books in English on the literature of Ceylon (Sri Lanka), prominent mention can be made of CHARLES EDMUND GODAKUMBURA, *The Literature of Ceylon* (1963).

Music: ARNOLD A. BAKE, "The Music of India," in *The New Oxford History of Music*, vol. 1 (1957), a general chapter on Indian music dealing with the philosophical background, Vedic chant, the ancient musical system, the modern classical system, and musical instruments; ELISE B. BARNETT, "Special Bibliography: The Art Music of India," *Ethnomusicology*, 14:278–312 (1970), a listing of books and articles on Indian music, published since 1959, which includes some publications on folk and religious music, dance, and drama; SUDHIBHUSHAN BHATTACHARYA, *Ethnomusicology and India* (1968), a synchronic approach attempting to relate folk, tribal, religious, and classical music in terms of the cultural background—includes outline notations in Indian *sargam;* ALAIN DANIELOU, *The Rāga-s of Northern Indian Music* (1968), an individualistic interpretation of modern North Indian *ragas*, based partly on ancient theory; B.C. DEVA, *Psychoacoustics of Music and Speech* (1967), a collection of articles on various aspects of music and speech, with emphasis on acoustics and the scientific study of Indian music; ARTHUR H. FOX-STRANGWAYS, *The Music of Hindostan* (1914, reprinted 1965), a work of wide scope including discussion of Vedic chant, folk music, and modern Indian classical music (includes notations in Western staff and analogies with Western music); O.C. GANGOLY, *Rāgas and Rāginīs*, 2 vol. (1934–35, reprinted 1948), a historical study that traces the systems of classifying *ragas* in Indian musical treatises, including a discussion of the time-theory of *ragas* as well as their iconography; NAZIR A. JAIRAZBHOY, *The Rāgas of North Indian Music* (1971), a technical work dealing with the structure and evolution of North Indian *ragas* (includes a 45 r.p.m. record of *raga* demonstrations performed on the *sitar* by VILAYAT KHAN, and extensive notations in Western staff and Indian *sargam*); BABURAO JOSHI and A. LOBO, *Introducing Indian Music* (n.d.), a series of four long-playing records, including musical examples and commentary, illustrating the main features of North Indian classical music, with accompanying booklet; WALTER KAUFMANN, *The Rāgas of North India* (1968), description and notations, in Western staff, of about 230 *ragas* of modern North Indian music; ALLEN KEESE, *The Sitar Book* (1968), an elementary guide to the *sitar*, with some description of playing techniques, musical exercises, and *gats* in ten *ragas;* HERBERT A. POPLEY, *The Music of India*, 3rd ed. (1970), a general work including discussion of historical background, scale, *raga, tala*, musical form, and instruments, with notations in Western staff and Indian *sargam;* HAROLD S. POWERS, "Indian Music and the English Language: A Review Essay," *Ethnomusicology*, 9:1–12 (1965), a survey of the most important literature on Indian music written in the English language; and "An Historical and Comparative Approach to the Classification of Rāgas (with an Appendix on Ancient Indian Tunings)," in *Selected Reports* of the Instiute of Ethnomusicology, UCLA, 1:1–78 (1970), a scholarly monograph that attempts to show the relationship between North and South Indian *ragas* that bear the same name but now differ in scale—includes material on ancient Indian music and gives many musical examples in Western staff; SWAMI PRAJNANANDA, *A History of Indian Music*, vol. 1 (1963), a technical work dealing with the origins and the music of the ancient period; P. SAMBAMOORTHY, *South Indian Music*, 5 vol. (1958–69), a comprehensive work covering many aspects of South Indian musical theory, both synchronically and diachronically; RAVI SHANKAR, *My Music, My Life* (1969), a general book combining autobiographical and biographical material with a discussion of Indian music history, theory, and instruments, including a manual for the *sitar*. BONNIE C. WADE, *Music in India: The Classical Traditions* (1979), a discussion of performance, theory, and basic instruments of both the North and South; M.R. GUATAM, *The Musical Heritage of India* (1981), a history with emphasis on classical traditions of both the North and South; WIM VAN DER MEER, *Hindustan: Music in the Twentieth Century* (1980), an introduction with emphasis on vocal music; DANIEL M. NEUMAN, *The Life of Music in North India* (1980), an analysis of the place in society of the musician.

Dance and theatre: The main source book of Indian classical dance and theatre is the *Nātya-śāstra*, ascribed to BHARATA MUNI, trans. by MANMOHAN GHOSH, 2 vol. (1950–61), which deals with ceremonies, gesture language, architecture, production styles, makeup, and costumes. A.K. COOMARASWAMY, *The Dance of Shiva*, rev. ed. (1957), brings alive the philosophy and aesthetics of Hindu dance. For general understanding of classical dance forms and techniques, see RINA SINGHA and REGINALD MASSEY, *Indian Dances* (1967), which includes semiclassical styles, modern ballet, and biographical notes on dancers and gurus. FAUBION BOWERS, *The Dance in India* (1953), is a fascinating description of the four major classical dances. K. BHARATHA IYER, *Kathakali* (1955), is perhaps the best work on this dance-drama, with detailed descriptions of characters, historical background, and interpretation of dramatic symbols, with photographs and line drawings. KAPILA VATSYAYAN, *Classical Indian Dance in Literature and the Arts* (1968), surveys dance as found in temple sculpture, the plastic arts, and literature—a scholarly work with photos. For tribal dances and ceremonies of Central India, see VERRIER ELWIN, *The Muria and Their Ghotul* (1947; abridged ed., *The Kingdom of the Young*, 1968). For dramatic forms, production techniques, and classical rules the best source book is also the *Nātya-śāstra*, ascribed to BHARATA MUNI, *op. cit.* ARTHUR B. KEITH, *Sanskrit Drama in Its Origin, Development, Theory and Practice* (1924, reprinted 1959), remains a standard scholarly critical work for comparative analysis of Sanskrit plays. P. LAL, *Great Sanskrit Plays* (1964) are actable transcreations in crisp modern English. A delightfully-written survey is FAUBION BOWERS, *Theatre in the East* (1956), which puts Indian dance and theatre in the larger perspective of South Asia with vivid and perceptive descriptions. For a general survey of classical, folk, and modern drama with sidelights on opera and ballet, illustrated by photographs and sketches, see BALWANT GARGI, *Theatre in India* (1962). HEMENDRA NATH DAS GUPTA, *The Indian Stage*, 4 vol. (1934–44), deals mainly with the growth of Bengali theatres, actors, and productions in exhaustive detail with reproductions of period notebooks and papers. BALWANT GARGI, *Folk Theatre in India* (1966), is an eyewitness account of religious, secular, and masked dramas in villages, with over 100 photographs and line sketches. See also J.C. MATHUR, *Drama in Rural India* (1964). BERYL DE ZOETE, *Dance and Magic Drama in Ceylon* (1957), is a vivid account of rituals and magical masked dances in a diary form of day-to-day performances. E.R. SARACHCHANDRA, *The Folk Drama of Ceylon*, 2nd ed. (1966), is a scholarly work on the development and background of Sinhalese folk dramas and cults of exorcism. For puppets and masks, see two small pamphlets: J. TILAKASIR, "Puppetry in Ceylon" and SIRI GUNASINGHE, "Masks of Ceylon" (both published by the Department of Cultural Affairs, Ceylon, 1962). For the history of Kolam and description of various characters see O. PERTOLD, "The Ceremonial Dances of the Sinhalese." *Archiv Orientální*, vol. 2 (1930). For Devil Dances and their interpretation in the light of witchcraft rituals, see DANDRIS DE SILVA GUNARATNA, "Demonology and Witchcraft in Ceylon," *J. Ceylon Brch. R. Asiat. Soc.*, vol. 4, no. 13 (1861); PAUL WIRZ, *Exorzismus und Heilkunde auf Ceylon* (1941); RACHEL VAN M. BAUMER and JAMES R. BRANDON (eds.), *Sanskrit Drama in Performance* (1981), essays on Ancient Indian theatrical performance.

Visual arts (General works): VINCENT A. SMITH, *A History of Fine Art in India and Ceylon*, 3rd ed. rev. and enl. (1962); A.K. COOMARASWAMY, *History of Indian and Indonesian Art* (1927, reprinted 1965); and BENJAMIN ROWLAND, *Art and Architecture of India*, 3rd ed. rev. (1967), are general introductions with good bibliographies. A.K. COOMARASWAMY, *Figures of Speech or Figures of Thought* (1946), and *Tranformation of Nature in Art* (1934, reprinted 1956), contain important essays which discuss Indian aesthetic theories from the traditional point of view. A clear classification of Hindu images with particular reference to south India has been made in T.A. GOPINATHA RAO, *Elements of Hindu Iconography*, 2 vol. in 4 (1914–16, reprinted 1968); and J.N. BANERJEA, *The Development of Hindu Iconography*, 2nd ed. rev. and enl. (1956), is an analytical introduction to the subject. N.K. BHATTASALI, *Iconography of Buddhist and Brahmanical Sculptures*

in the Dacca Museum (1929); BENOYTOSH BHATTACHARYYA, *The Indian Buddhist Iconography*, 2nd ed. (1958); and ALFRED FOUCHER, *Étude sur l'iconographie bouddhique de l'Inde*, 2 vol. (1900–05), are important studies of Buddhist iconography. A.K. COOMARASWAMY, *Yaksas*, 2 vol. (1928–31), is a masterly study of early iconography. HEINRICH ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946, reprinted 1963), discusses some important symbols of Indian art. Good collections of photographs are in JAMES BURGESS, *The Ancient Monuments, Temples and Sculptures of India*, 2 vol. (1897–1911); A.K. COOMARASWAMY, *Viśvakarma* (1912); STELLA KRAMRISCH, *The Art of India* (1954); and HEINRICH ZIMMER, *The Art of Indian Asia*, 2 vol. (1955).

Architecture: JAMES FERGUSSON, *History of Indian and Eastern Architecture*, rev. ed., 2 vol. (1910, reprinted 1967); PERCY BROWN, *Indian Architecture*, 2 vol., 5th ed. (1965); and S.K. SARASWATI, "Architecture," in R.C. MASUMDAR (ed.), *History and Culture of the Indian People*, vol. 3 and 5 (1954–57), are standard works that survey the entire history of Indian architecture. JAMES FERGUSSON and JAMES BURGESS, *The Cave Temples of India* (1880, reprinted 1969), is a comprehensive account of rock-cut architecture. STELLA KRAMRISCH, *The Hindu Temple*, 2 vol. (1946), is concerned with principles and symbolism, and KRISHNA DEVA, *Temples of North India* (1969), presents a synoptic view of the various north Indian styles. G. JOUVEAU-DUBREUIL, *Archéologie du sud de l'Inde*, 2 vol. (1914), analyses the south Indian style. K.R. SRINVASAN, *Cave Temples of the Pallavas* (1964); and ARTHUR H. LONGHURST, *Pallava Architecture*, 3 vol. (1924–30), are important studies of early south Indian architecture. SIR JOHN MARSHALL, "The Monuments of Muslim India," in *The Cambridge History of India*, vol. 3, pp. 568–640 (1928, reprinted 1965); and PERCY BROWN, "Monuments of the Mughal Period," *ibid.*, vol. 4, pp. 523–576 (1937, reprinted 1963), are important essays on Islamic architecture in India. See also ELIZABETH S. MERKLINGER, *Indian Islamic Architecture: The Deccan 1347–1686* (1981).

Sculpture: S.K. SARASVATI, *A Survey of Indian Sculpture* (1957); and STELLA KRAMRISCH, *Indian Sculpture* (1933), discuss the broad historical and stylistic trends. LUDWIG BACHHOFER, *Early Indian Sculpture* (1929), is a fine stylistic analysis of sculptures from the third century B.C. to the third century A.D. The classic work on Gandhāra art is ALFRED FOUCHER, *L'Art gréco-bouddhique du Gandhâra*, 2 vol. (1905–41), though several of its conclusions are no longer tenable. SIR JOHN MARSHALL, *Taxila*, 3 vol. (1951), discusses works recovered from that site; and HAROLD INGHOLT, *Gandharan Art in Pakistan* (1957, reprinted 1971), provides excellent photographic documentation. R.D. BANERJI, *The Age of the Imperial Guptas* (1933), has a general discussion of the Gupta style; D.R. SAHNI, *Catalogue of the Museum of Archaeology at Sārnāth* (1914), contains information of interest on the school of Sārnāth. ELIKY ZANNAS, *Khajurāho* (1960); K.C. PANIGRAHI, *Archaeological Remains at Bhubaneswar* (1961); and R.D. BANERJI, *Eastern Indian School of Medieval Sculpture* (1933), are monographs on schools of north Indian medieval sculpture. A.H. LONGHURST, *op. cit.*; and K.A. NILAKANTA SASTRI, *The Cōlas*, 2nd ed. rev. (1955), contain information on the south Indian medieval styles. C. SIVARAMAMURTI, *South Indian Bronzes* (1963); P.R. SRINIVASAN, *Bronzes of South India* (1963); DOUGLAS E. BARRETT, *Early Cōla Bronzes* (1965), are important studies of south Indian bronzes; FREDERICK M. ASHER, *The Art of Eastern India: 300–800* (1980), a study of the pre-Pala period.

Painting: In considering the published literature it is important to remember that the study of Indian painting, confined to a limited number of scholars, is of comparatively recent growth, and is therefore full of controversies and uncertainties which keep shifting with the discovery of fresh materials. The standard work on Ajanta is GHULAM YAZDANI, *Ajanta*, 4 vol. (1930–55); and on the western Indian style, MOTI CHANDRA, *Jain Miniature Paintings from Western India* (1949). Much interesting information on the period of transition to the Rajasthani and Mughal styles has been brought together in KARL J. KHANDALAVALA and MOTI CHANDRA, *New Documents of Indian Painting* (1969). The Mughal school has been ably presented in PERCY BROWN, *Indian Painting under the Mughals, A.D. 1550 to A.D. 1750* (1924); and STUART C. WELCH, *The Art of Mughal India* (1964). DOUGLAS E. BARRETT, *Painting of the Deccan, XVI–XVII Century* (1958), is a brief introduction to the subject. The main publication on the Company style is MILDRED and WILLIAM G. ARCHER, *Indian Painting for the British, 1770–1880* (1955). The classic work on Pahari and Rajasthani painting is A.K. COOMARASWAMY's pioneering *Rajput Painting*, 2 vol. (1916). Fresh discoveries which have considerably changed the understanding of its history are summarized in KARL KHANDALAVALA, MOTI CHANDRA, and PRAMOD CHANDRA, *Miniature Painting: A Catalogue of the Exhibition of the Sri Motichand Khajanchi Collection* (1960). MOTI CHANDRA, *Mewar Painting in the Seventeenth Century* (1957); WILLIAM G. ARCHER, *Indian Painting in Bundi and Kotah* (1959); PRAMOD CHANDRA, *Bundi Painting* (1959); ERIC DICKINSON and KARL KHANDALAVALA, *Kishangarh Painting* (1959); WILLIAM G. ARCHER, *Central Indian Painting* (1958); and ANAND KRISHNA, *Malwa Painting* (1963), are informative summaries of the growing knowledge of the various schools of Rājasthān. The standard work, illustrated copiously, on the Pahari style is KARL KHANDALAVALA, *Pahārī Miniature Painting* (1958), and different in its account from WILLIAM G. ARCHER, *Indian Painting in the Punjab Hills* (1952). Books which cover most of the schools and are profusely illustrated include N.C. MEHTA, *Studies in Indian Painting* (1926); WILLIAM G. ARCHER, *Indian Miniatures* (1960); ROBERT SKELTON, *Indian Miniatures from the XVth to the XIXth Centuries* (1961); DOUGLAS E. BARRETT and BASIL GRAY, *Painting of India* (1963); and STUART C. WELCH and MILO C. BEACH, *Gods, Thrones and Peacocks: Northern Indian Painting from Two Traditions, Fifteenth to Nineteenth Centuries* (1965). (*Ceylon*): Standard works on Sinhalese art are SENERAT PARANAVITANA, *The Stūpa in Ceylon* (1946), *Art and Architecture of Ceylon: Polonnaruva Period* (1954), and *Ceylon: Paintings from Temple, Shrine and Rock* (1957). MOTI CHANDRA, *Studies in Early Indian Painting* (1975), covers the 5th through 16th centuries; CALAMBUR SIVARAMAMURTI, *The Art of India* (1977), covering all types of art with 1175 illustrations; STUART C. WELCH, *Room for Wonder: Indian Painting During the British Period, 1760–1880* (1978), an exhibition catolog with detailed comments on 125 illustrations.

Decorative arts: Scholarly literature on the decorative arts in India is scanty and mainly in learned journals. The *Journal of Indian Art and Industries* (1886–1916) is the most important and contains numerous pioneering studies. SIR GEORGE WATT, *Indian Art at Delhi* (1903); and SIR GEORGE BIRDWOOD, *The Industrial Arts of India*, 2 vol. (1880), are for the most part descriptive texts emphasizing the technical aspects of the decorative arts as they had survived up to the closing years of the 19th century. JOHN IRWIN, "Textiles and the Minor Arts," in LEIGH ASHTON (ed.), *The Art of India and Pakistan*, pp. 201–237 (1950), is a brief historical survey of the subject. MOTI CHANDRA, "Ancient Indian Ivories," *Bulletin of the Prince of Wales Museum of Western India*, no. 6, pp. 4–63 (1957–59), is a monograph on the history of Indian ivory carving. THOMAS H. HENDLEY, *Asian Carpets* (1905), treats Indian examples in the important collections of the Maharaja of Jaipur. GEORGE P. BAKER, *Calico Painting and Printing in the East Indies in the XVIIth and XVIIIth Centuries* (1921), is the most important work on the subject; WILBRAHAM EGERTON, *An Illustrated Handbook of Indian Arms* (1880), catalogs and describes the wide range and achievement of the armourer's craft.

(C.S./J.A.B.v.B./E.C.D./C.M.N./A.K.R./N.A.J./B.Ga./P.Ch.)

Southeast Asia

Southeast Asia is the vast region of Asia situated east of the Indian subcontinent and south of China. It consists of two dissimilar portions: a continental projection (commonly called mainland Southeast Asia) and a string of archipelagoes to the south and east of the mainland (insular Southeast Asia). Extending some 700 miles (1,100 kilometres) southward from the mainland into insular Southeast Asia is the Malay Peninsula; this peninsula structurally is part of the mainland, but it also shares many ecological and cultural affinities with the surrounding islands and thus functions as a bridge between the two regions.

Mainland Southeast Asia is divided into the countries of Cambodia, Laos, Myanmar (Burma), Thailand, Vietnam, and the small city-state of Singapore at the southern tip of the Malay Peninsula; Cambodia, Laos, and Vietnam, which occupy the eastern portion of the mainland, often are collectively called the Indochinese Peninsula. Malaysia is both mainland and insular, with a western portion on the Malay Peninsula and an eastern part on the island of Borneo. Except for the small sultanate of Brunei (also on Borneo), the remainder of insular Southeast Asia consists of the archipelagic nations of Indonesia and the Philippines.

Southeast Asia stretches some 4,000 miles at its greatest extent (roughly from northwest to southeast) and encompasses some 5,000,000 square miles (13,000,000 square kilometres) of land and sea, of which about 1,736,000 square miles is land. Mount Hkakabo in northern Myanmar on the border with China, at 19,295 feet (5,881 metres), is the highest peak of mainland Southeast Asia. Although the modern nations of the region are sometimes thought of as being small, they are—with the exceptions of Singapore and Brunei—comparatively large. Indonesia, for example, is more than 3,000 miles from west to east (exceeding the west-east extent of the continental United States) and more than 1,000 miles from north to south; the area of Laos is only slightly smaller than that of the United Kingdom; and Myanmar is considerably larger than France.

All of Southeast Asia falls within the tropical and subtropical climatic zones, and much of it receives considerable annual precipitation. It is subject to an extensive and regular monsoonal weather system (*i.e.*, one in which the prevailing winds reverse direction every six months) that produces marked wet and dry periods in most of the region. Southeast Asia's landscape is characterized by three intermingled physical elements: mountain ranges, plains and plateaus, and water in the form of both shallow seas and extensive drainage systems. Of these, the rivers probably have been of the greatest historical and cultural significance, for waterways have decisively shaped forms of settlement and agriculture, determined fundamental political and economic patterns, and helped define the nature of Southeast Asians' worldview and distinctive cultural syncretism. It also has been of great importance that Southeast Asia, which is the most easily accessible tropical region in the world, lies strategically astride the sea passage between East Asia and the Middle Eastern–Mediterranean world.

Within this broad outline, Southeast Asia is perhaps the most diverse region on Earth. The number of large and small ecological niches is more than matched by a staggering variety of economic, social, and cultural niches Southeast Asians have developed for themselves; hundreds of ethnic groups and languages have been identified. Under these circumstances, it often is difficult to keep in mind the region's underlying unity, and it is understandable that Southeast Asia should so often be treated as a miscellaneous collection of cultures that simply do not quite fit anywhere else.

Yet from ancient times Southeast Asia has been considered by its neighbours to be a region in its own right and not merely an extension of their own lands. The Chinese called it Nanyang and the Japanese Nan'yō, both names meaning "South Seas," and South Asians used such terms as Suvarnabhūmi (Sanskrit: "Land of Gold") to describe the area.

Modern scholarship increasingly has yielded evidence of broad commonalities uniting the peoples of the region across time. Studies in historical linguistics, for example, have suggested that the vast majority of Southeast Asian languages—even many of those previously considered to have separate origins—either sprang from common roots or have been long and inseparably intertwined. Despite inevitable variation among societies, common views of gender, family structure, and social hierarchy and mobility may be discerned throughout mainland and insular Southeast Asia, and a broadly common commercial and cultural inheritance has continued to affect the entire region for several millennia. These and other commonalities have yet to produce a conscious or precise Southeast Asian identity, but they have given substance to the idea of Southeast Asia as a definable world region and have provided a framework for the comparative study of its components. (W.H.F.)

This article first provides a general overview of the Southeast Asian region, followed by a detailed treatment of Brunei, Cambodia, Laos, Malaysia, Myanmar, Singapore, Thailand, and Vietnam. Although Indonesia and the Philippines are discussed below in their regional context, the detailed treatments of these two countries are found, respectively, in the articles INDONESIA and PHILIPPINES.

The article is divided into the following sections:

The region 712
  Physical and human geography 712
    The land
    The people
    The economy
  History 722
    Early society and accomplishments
    The classical period
    Patterns of a colonial age
    Contemporary Southeast Asia
The countries of Southeast Asia 729
  Brunei 729
    Physical and human geography
    History
  Cambodia 730
    Physical and human geography
    History
  Laos 740
    Physical and human geography
    History
  Malaysia 745
    Physical and human geography
    History
  Myanmar 755
    Physical and human geography
    History
  Singapore 765
    Physical and human geography
    History
  Thailand 769
    Physical and human geography
    History
  Vietnam 779
    Physical and human geography
    History
Bibliography 791

THE REGION

Physical and human geography

THE LAND

Geology and relief. The physiography of Southeast Asia has been formed to a large extent by the convergence of three of the Earth's major crustal units: the Eurasian, Indian-Australian, and Pacific plates. The land has been subjected to a considerable amount of faulting, folding, uplifting, and volcanic activity over geologic time, and much of the region is mountainous. There are marked structural differences between the mainland and insular portions of the region.

Mountain ranges

Mainland Southeast Asia. The mainland is characterized by a series of generally north–south-trending mountain ranges separated by a number of major river valleys and their associated deltas. In many ways these ranges resemble ribs in a fan, where the interstices are deep trenches carved by the rivers. Although the mainland as a whole is similar in a structural sense, its various geologic components and the time periods of their orogenic (mountain-building) episodes differ. Much of the region has been affected by the gradual, continuing collision of the Indian subcontinent with the Eurasian Plate over roughly the past 50 million years, an event that—with diminishing intensity from west to east—has been responsible for deforming the land. Nonetheless, mainland Southeast Asia is relatively stable geologically, with no active or recently active volcanoes and, except in the northwest and north, little seismic activity.

The ranges fan out southward from the southeastern corner of the Plateau of Tibet, where they are tightly spaced. A major rib of this system extends through the entire western margin of Myanmar (Burma); describing an elongated letter S, it consists of (from north to south) the Pātkai Range, Nāga Hills, Chin Hills, and Arakan Mountains. Farther to the south the same rib emerges from beneath the sea to become the Andaman and Nicobar Islands of India.

Another major system extends along a straight north-south axis from eastern Myanmar east of the Salween River through northwestern Thailand to south of the Isthmus of Kra on the Malay Peninsula. It consists of a series of elongated blocks rather than one continuous ridge. The core of these blocks is granite, which has intruded into previously folded and faulted limestone and sandstone. The altitudes of the ranges diminish from above 8,000 feet (2,440 metres) on the Chinese border in the north to below 4,000 feet on the Isthmus of Kra, and the ranges are spread farther apart toward the south.

John Elk/Tony Stone Images

Karst-limestone landscape of the Malay Peninsula along the Krabi River, southern Thailand.

The easternmost major mountain feature on the mainland is the Annamese Cordillera (Chaîne Annamitique) in Laos and Vietnam. In the portion between Laos and Vietnam, the chain forms a nearly straight spine of ranges from northwest to southeast, with a steep face rising from the South China Sea to the east and a more gradual slope to the west. The mountains thin out considerably south of Laos and become asymmetrical in form. The upland zone is characterized by a number of plateau remnants.

The rather neat fanlike pattern of the mountain ranges is interrupted occasionally by several old blocks of strata that have been folded, faulted, and deeply dissected. These ancient massifs now form either low platforms or high plateaus. The westernmost of these, the Shan Plateau of eastern Myanmar, measures some 250 miles (400 kilometres) from north to south and 75 miles from east to west and has an average elevation of about 3,000 feet. The largest of these features is the Korat Plateau in eastern Thailand and west-central Laos. This area actually is more of a low platform, which on average is only a few hundred feet above the floodplains of the surrounding rivers. It consists of a string of hills that direct surface drainage eastward to the Mekong River. The hills range in elevation from 500 to 2,000 feet, with the highest altitudes occurring near the southwestern rim.

Plateaus

The broad river valleys between the uplands and the even wider deltas at the southernmost points contain most of the mainland's lowland areas. These regions generally are covered with alluvial sediments that support much of the mainland's cultivation and, in turn, most of its population centres. The most extensive coastal lowland is the lower Mekong basin, which encompasses most of Cambodia and southern Vietnam. The Cambodian portion is a broad, bowl-shaped area lying just above sea level, with numerous hill outcrops jutting above the landscape; at its centre is a large freshwater lake, the Tonle Sap. To the south the river's vast, flat delta occupies the entire southern tip of Vietnam. Outside the river deltas, the coastal lowlands are little more than narrow strips between the mountains and the sea, except around the southern half of the Malay Peninsula.

The Malay Peninsula stretches south for some 900 miles from the head of the Gulf of Thailand (Siam) to Singapore and thus extends the mainland into insular Southeast Asia. The narrowest point, the Isthmus of Kra (about 40 miles wide), also roughly divides the peninsula into two parts: the long linear mountain ranges of the northern part described above give way just south of the isthmus to blocks of short, parallel ranges aligned north-south, so that the southern portion trends to the southeast and becomes much wider. In areas such as the west coast between southern Thailand and northwestern Malaysia, distinctive karst-limestone landscapes have developed. Peaks on the peninsula range from 5,000 to 7,000 feet in elevation.

Insular Southeast Asia. Characteristic of insular (or archipelagic) Southeast Asia are the chains of islands—the Malay and Philippine archipelagoes—that have been formed along the boundaries of the three crustal segments of the Earth that meet there. Crustal instability is marked throughout the region. Earthquakes and volcanic activity are quite common along the entire southern and eastern margin. One consequence of the seismic activity is that a large number of lakes are found in the region.

The Sunda Shelf

Dominating the region is the Sunda Shelf, the portion of the Asian continental shelf that extends southward from the Gulf of Thailand to the Java Sea. Where the shelf meets and overrides the oceanic crust to the south, the vast volcanic arc of the Greater and Lesser Sunda islands have been formed. The islands are characterized by highland cores, from which flow short rivers across the narrow coastal plains. The shallow waters of the Sunda Shelf are as important to the inhabitants as the land, since the sea has facilitated communication and trade among the islands. At one time, sea levels were considerably lower than now, and land bridges existed on the Sunda Shelf

that connected the islands and allowed plants and animals to migrate throughout the region.

The extreme southeastern islands of Southeast Asia—the eastern Moluccas (Maluku) and the island of New Guinea—lie on the Sahul Shelf, a northwestern extension of Australia, and structurally are not part of Asia. In the east the Philippine Islands rise between two blocks of sinking (subducted) oceanic crust at the boundary of the Eurasian and Pacific plates.

Drainage. Mainland Southeast Asia is drained by five major river systems, which from west to east are the Irrawaddy, Salween, Chao Phraya, Mekong, and Red rivers. The three largest systems—the Irrawaddy, Salween, and Mekong—have their origins in the Plateau of Tibet. These three rivers are somewhat atypical: their middle and upper drainage basins are not broad catchment areas with many small tributaries feeding larger ones but rather consist of a few streams confined to narrow, closely spaced valleys.

The Irrawaddy River flows through western Myanmar, draining the eastern slope of the country's western mountain chain and the western slope of the Shan Plateau. Although the river itself is shorter than either the Salween or the Mekong rivers, its lowland areas are more extensive. Most conspicuous is its delta, which is about 120 miles wide at its base and is expanding rapidly into the Andaman Sea.

The Salween River flows for several hundred miles through southern China before entering eastern Myanmar. In contrast to the Irrawaddy, the Salween is a highlands river throughout nearly all of its course. Its drainage basin is highly restricted with few tributaries, and its delta area is small. Even though the Salween's catchment area is limited and is sheltered from seasonal rains, its water volume fluctuates considerably from season to season.

The Mekong basin

The Mekong—one of the world's great river systems—is the longest river of mainland Southeast Asia and has the largest drainage basin. After flowing for some 1,200 miles through southern China, the Mekong flows for nearly 1,500 more miles through Laos (where it also forms much of the western border of the country), Cambodia, and Vietnam. The Tonle Sap in Cambodia, the largest lake in Southeast Asia, drains into the vast Mekong delta. The area of the lake varies greatly with the precipitation cycle of the region.

The Chao Phraya River is the major river of Thailand and the shortest of the great rivers of the mainland. Rising in the northwestern highlands of Thailand, it drains the western portion of northern Thailand. The densely populated delta contains Bangkok, Thailand's capital and the largest city on the mainland. The Red River of northern Vietnam has the smallest drainage basin of the major rivers. The river follows a narrow valley through southern China and northwestern Vietnam before flowing into a relatively small lowland.

Soils. Southeast Asia, on balance, has a higher proportion of relatively fertile soils than most tropical regions, and soil erosion is less severe than elsewhere. Much of the region, however, is covered by tropical soils that generally are quite poor in nutrients. Often the profusion of plant life is more related to heat and moisture than to soil quality, even though these climatic conditions intensify both chemical weathering and the rate of bacterial action that usually improve soil fertility. Once the vegetation cover is removed, the supply of humus quickly disappears. In addition, the often heavy rainfall leaches the soils of their soluble nutrients, hastens erosion, and damages the soil texture. The leaching process in part results in laterites of reddish clay that contain hydroxides of iron and alumina.

Laterite soils are common in parts of Myanmar, Thailand, and Vietnam and also occur in the islands of the Sunda Shelf, notably Borneo. The most fertile soils occur in regions of volcanic activity, where the ejecta is chemically alkaline or neutral. Such soils are found in parts of Sumatra and much of Java in Indonesia. The alluvial soils of the river valleys also are highly fertile and are intensively cultivated.

Climate. All of Southeast Asia falls within the warm, humid tropics, and its climate generally can be characterized as monsoonal (*i.e.*, marked by wet and dry periods). Changing seasons are more associated with rainfall than with temperature variations. There is, however, a high degree of climatic complexity within the region.

Temperatures. Regional temperatures at or near sea level remain fairly constant throughout the year, although monthly averages tend to vary more with increasing latitude. Thus, with the exception of northern Vietnam, annual average temperatures are close to 80° F (27° C). Increasing elevation acts to decrease average temperatures, and such locations as the Cameron Highlands in peninsular Malaysia and Baguio in the Philippines have become popular tourist destinations in part because of their relatively cooler climates. Proximity to the sea also tends to moderate temperatures.

Precipitation. Much of Southeast Asia receives more than 60 inches (1,500 millimetres) of rainfall annually, and many areas commonly receive double and even triple that amount. The rainfall pattern is distinctly affected by two prevailing air currents: the northeast (or dry) monsoon and the southwest (or wet) monsoon.

The two monsoons

The northeast monsoon occurs roughly from November to March and brings relatively dry, cool air and little precipitation to the mainland. As the southwestward-flowing air passes over the warmer sea, it gradually warms and gathers moisture. Precipitation is especially heavy where the airstream is forced to rise over mountains or encounters a landmass. The east coast of peninsular Malaysia, the Philippines, and parts of eastern Indonesia receive the heaviest rains during this period.

The southwest monsoon prevails from May to September, when the air current reverses and the dominant flow is to the northeast. The mainland receives the bulk of its rainfall during this period. Over much of the southern Malay Peninsula and insular Southeast Asia there is little or no prolonged dry season. This is especially marked in much of the equatorial region and along the east coast of the Philippines.

While the dry and wet monsoons are important in explaining rainfall patterns, so too are such factors as relief, land and sea breezes, convectional overturning and cyclonic disturbances. These factors often are combined with monsoonal effects to produce highly variable rainfall patterns over relatively short distances. While many of the cyclonic disturbances produce only moderate rainfall, others mature into tropical storms—called cyclones in the Indian Ocean and typhoons in the Pacific—that bring heavy rains and destruction to the areas over which they pass. The Philippines are particularly affected by these storms.

Plant life. The seasonal nature and pattern of Southeast Asia's rainfall, as well as the region's physiography, have strongly affected the development of natural vegetation. The hot, humid climate and enormous variety of habitats have given rise to an abundance and diversity of vegetative forms unlike that in any other area of the world. Much of the natural vegetation has been modified by human action, although large areas of relatively untouched land still can be found.

The vegetation can be grouped into two broad categories: the tropical-evergreen forests of the equatorial lowlands and the open type of tropical-deciduous, or "monsoon," forests in areas of seasonal drought. The evergreen forests are characterized by multiple stories of vegetation, consisting of a variety of trees and plants. Although a large diversity of tree species is found in these forests, members of the Dipterocarpaceae family account for roughly half of the varieties. Deciduous forests are found in eastern Indonesia and those parts of the mainland where annual rainfall does not exceed 80 inches. Just as in the equatorial forest, a wide variety of species is normally the rule. Certain species, such as teak, have become highly valued commercially. Teak is found in parts of Indonesia, Myanmar, Thailand, and Laos.

In addition to these two basic types of vegetation, other regional patterns reflect topography. Especially noteworthy are coastal and highland plant communities. Mangrove belts, of which there are more than 30 varieties, occur where silt is deposited in coastal areas. Upland forests dominated by maples, oaks, and magnolias are found especially on mainland mountain slopes.


SINGAPORE
Scale 1:702,000
1 inch equals approx. 11 miles
Cassini Projection
MALAYSIA
Johor Baharu
SEMBAWANG
Yishun
Kranji War Memorial
SINGAPORE ISLAND
SERANGOON
CENTRAL CATCHMENT NATURE RESERVE
BUKIT PANJANG
JURONG
Nanyang University
National Univ of Singapore
PASIR PANJANG
Queenstown
Toa Payoh
CITY
KATONG
Bedok
CHANGI
Changi International Airport
Keppel Harbour
SENTOSA ISLAND
BUKUM ISLAND
SUDONG ISLAND
SAKRA ISLAND
AYER CHAWAN ISLAND
Singapore Strait
INDONESIA
Serimbun Reservoir
CHANGI Neighbourhoods
Yishun New towns
Major roads
Other roads
Railroads
Dams
Points of interest
National parks
Built-up areas
Scale 1:12,898,000
1 inch equals approx. 204 miles
Oblique Mercator Projection
CHINA
K'un-ming
Kuei-yang
Tropic of Cancer
INDIA
BHUTAN
BANGLADESH
Chittagong
MYANMAR
Mandalay
Yangon
Moulmein
Pegu
Bassein
Sittwe
KUMON RANGE
PATKAI RANGE
NAGA HILLS
MANIPUR HILLS
MANGIN RANGE
CHIN HILLS
ARAKAN MOUNTAINS
PEGU MOUNTAINS
SHAN PLATEAU
Taunggyi
DAWNA RANGE
BILAUKTAUNG RANGE
Gulf of Martaban
MERGUI ARCHIPELAGO
Isthmus of Kra
ANDAMAN SEA
ANDAMAN ISLANDS (India)
Bay of Bengal
LAOS
Vientiane
Louangphrabang
PLAIN OF JARS
THAILAND
Bangkok
Chiang Mai
Nakhon Ratchasima
KHORAT PLATEAU
Gulf of Thailand
CAMBODIA
Phnom Penh
VIETNAM
Hanoi
Haiphong
Gulf of Tonkin
Hue
Da Nang
ANNAMESE CORDILLERA
CENTRAL HIGHLANDS
Ho Chi Minh City
Can Tho
Nha Trang
Qui Nhon
Da Lat
PHU QUOC ISLAND
PARACEL ISLANDS (Administered by China; claimed by Vietnam)
SOUTH CHINA SEA
PHILIPPINES

MAP INDEX

Cities and towns

Akyab, see Sittwe
Allanmyo 19 22 N 95 13 E
Alor Setar 6 07 N 100 22 E
Amarapura 21 54 N 96 03 E
An Bien 9 45 N 105 00 E
An Khe 13 57 N 108 39 E
Ânlóng Vêng 14 14 N 104 05 E
Attapu 14 48 N 106 50 E
Ava 21 51 N 95 58 E
Ayutthaya, see Phra Nakhon Si Ayutthaya
Bâ Kêv 13 42 N 107 12 E
Ba Na 15 59 N 107 59 E
Bac Can 22 08 N 105 50 E
Bac Giang 21 16 N 106 12 E
Bac Lieu 9 17 N 105 43 E
Bac Ninh 21 11 N 106 03 E
Bago, see Pegu
Ban Houayxay 20 18 N 100 26 E
Ban Phai 16 04 N 102 44 E
Ban Phôntiou (Phon Tiou) 17 53 N 104 37 E
Ban Tha Kham (Tha Kham) 9 06 N 99 14 E
Bandar Maharani, see Muar
Bandar Seri Begawan (Brunei) 4 53 N 114 56 E
Bangar 4 43 N 115 04 E
Bangkok (Krung Thep) 13 45 N 100 31 E
Bao Loc 11 32 N 107 48 E
Bassein (Pathein) 16 47 N 94 44 E
Bătdâmbâng (Battambang) 13 06 N 103 12 E
Batu Pahat 1 51 N 102 56 E
Bau 1 25 N 110 09 E
Bedok, *new town* 1 19 N 103 57 E
Ben Thuy 18 39 N 105 42 E
Bentong 3 32 N 101 55 E
Betong 5 45 N 101 05 E
Bhamo 24 16 N 97 14 E
Bien Hoa 10 57 N 106 49 E
Bien Son (Bim Son) 20 04 N 105 51 E
Bintulu 3 10 N 113 02 E
Bogale 16 17 N 95 24 E
Bong Son 14 26 N 109 01 E
Brunei, see Bandar Seri Begawan
Bukit Mertajam 5 22 N 100 28 E
Buon Me Thuot (Lac Giao) 12 40 N 108 03 E
Buriram 15 00 N 103 07 E
Butterworth 5 25 N 100 24 E
Ca Mau 9 11 N 105 08 E
Cam Pha 21 01 N 107 19 E
Cam Ranh 11 54 N 109 13 E
Can Tho 10 02 N 105 47 E
Cao Bang 22 40 N 106 15 E
Cao Lanh 10 27 N 105 38 E
Cau Giat 19 09 N 105 38 E
Cha-am 12 48 N 99 58 E
Chaiyaphum 15 48 N 102 02 E
Champasak 14 53 N 105 52 E
Chanthaburi (Chantabun) 12 36 N 102 09 E
Chau Doc 10 42 N 105 07 E
Chauk 20 53 N 94 49 E
Chbar 12 46 N 107 10 E
Cheo Reo 13 24 N 108 27 E
Chiang Mai (Chiengmai) 18 47 N 98 59 E
Chiang Rai (Chiengrai) 19 54 N 99 50 E
Cho Don 22 11 N 105 39 E
Chŏâm Khsant 14 13 N 104 56 E
Chon Buri 13 22 N 100 59 E
Con Son 8 41 N 106 37 E
Chŏng Kal 13 57 N 103 35 E
Chum Phae 16 32 N 102 06 E
Cong Tum, see Kon Tum
Da Lat 11 56 N 108 25 E
Da Nang (Tourane) 16 04 N 108 13 E
Daik-U 17 47 N 96 40 E
Dawei, see Tavoy
Dien Chau 18 59 N 105 37 E
Dong Ha 16 49 N 107 08 E
Dong Hoi 17 29 N 106 36 E
Dungun, see Kuala Dungun
Gangaw 22 10 N 94 08 E
George Town (Pinang) 5 25 N 100 20 E
Gia Rai 9 14 N 105 28 E
Go Cong 10 22 N 106 40 E
Gwa 17 36 N 94 35 E
Ha Giang 22 50 N 104 59 E
Ha Tinh 18 20 N 105 54 E
Hai Duong 20 56 N 106 19 E
Haiphong (Hai Phong) 20 52 N 106 41 E
Haka (Hakha) 22 39 N 93 37 E
Hanoi (Ha Noi) 21 02 N 105 51 E
Hat Yai (Haad Yai) 7 01 N 100 28 E
Henzada 17 38 N 95 28 E
Ho Chi Minh City (Saigon) 10 45 N 106 40 E
Ho Xa 17 04 N 107 02 E
Hoa Binh 20 50 N 105 20 E
Hoi An 15 52 N 108 19 E
Homalin 24 52 N 94 55 E
Hong Gai (Hon Gai) 20 57 N 107 05 E
Hpa-an, see Pa-an
Hsipaw 22 37 N 97 18 E
Hua Hin 12 34 N 99 58 E
Hue 16 28 N 107 36 E
Hung Yen 20 39 N 106 04 E
Insein 16 53 N 96 07 E
Ipoh 4 35 N 101 05 E
Jelebu, see Kuala Kelawang
Jesselton, see Kota Kinabalu
Johor Baharu 1 28 N 103 45 E
Kajang 2 59 N 101 47 E
Kalasin 16 26 N 103 30 E
Kalaw 20 38 N 96 34 E
Kale 16 05 N 97 54 E
Kampar 4 18 N 101 09 E
Kamphaeng Phet 16 28 N 99 30 E
Kâmpóng Cham 12 00 N 105 27 E
Kâmpóng Chhnăng 12 15 N 104 40 E
Kâmpóng Kdei 13 07 N 104 21 E
Kampong Labi (Labi) 4 23 N 114 27 E
Kâmpóng Saôm (Sihanoukville) 10 38 N 103 30 E
Kâmpóng Spoe 11 27 N 104 32 E
Kâmpóng Thum 12 42 N 104 54 E
Kâmpôt 10 37 N 104 11 E
Kanchanaburi 14 01 N 99 32 E
Kangar 6 26 N 100 12 E
Karak 3 24 N 102 02 E
Katha 24 11 N 96 21 E
Kawkareik 16 33 N 98 14 E
Kawlin 23 47 N 95 41 E
Kawthaung 9 59 N 98 33 E
Kelang (Klang) 3 02 N 101 27 E
Kĕng Tung 21 17 N 99 36 E
Keluang 2 02 N 103 19 E
Khon Kaen 16 26 N 102 50 E
Khorat, see Nakhon Ratchasima
Kon Tum (Cong Tum or Kontun) 14 21 N 108 00 E
Kota Baharu 6 08 N 102 15 E
Kota Kinabalu (Jesselton) 5 59 N 116 04 E
Kota Tinggi 1 44 N 103 54 E
Krabi 8 04 N 98 55 E
Krâchéh (Kratie) 12 29 N 106 01 E
Krâkôr 12 32 N 104 12 E
Krŏng Kaôh Kŏng 11 37 N 102 59 E
Krung Thep, see Bangkok
Kuala Belait 4 35 N 114 11 E
Kuala Dungun (Dungun) 4 47 N 103 26 E
Kuala Kangsar 4 46 N 100 56 E
Kuala Kelawang (Jelebu) 2 56 N 102 05 E

Kuala Lipis ..... 4 11 N 102 03 E
Kuala Lumpur ... 3 10 N 101 42 E
Kuala Terengganu .... 5 20 N 103 08 E
Kuantan ..... 3 48 N 103 20 E
Kuching ..... 1 33 N 110 20 E
Kulai ..... 1 40 N 103 36 E
Kulim ..... 5 22 N 100 34 E
Kutkai ..... 23 27 N 97 56 E
Kyaikkami ..... 16 04 N 97 34 E
Kyaiklat ..... 16 26 N 95 44 E
Kyaikto ..... 17 18 N 97 01 E
Kyaukme ..... 22 32 N 97 02 E
Kyaukpadaung .. 20 50 N 95 08 E
Kyaukpyu (Ramree) ..... 19 05 N 93 52 E
Kyaukse ..... 21 36 N 96 08 E
Kyauktaga ..... 18 10 N 96 37 E
Labi, see Kampong Labi
Labuan, see Victoria
Labutta ..... 16 09 N 94 46 E
Lac Giao, see Buon Me Thuot
Lai Chau ..... 22 04 N 103 10 E
Lampang ..... 18 18 N 99 31 E
Lamphun (Lampoon) ..... 18 35 N 99 01 E
Lao Bao ..... 16 37 N 106 36 E
Lao Cai ..... 22 30 N 103 58 E
Lashio ..... 22 56 N 97 45 E
Lawksawk ..... 21 15 N 96 52 E
Le Thuy ..... 17 14 N 106 49 E
Letpadan ..... 17 47 N 95 45 E
Lewe ..... 19 38 N 96 07 E
Loei ..... 17 29 N 101 35 E
Loi-kaw ..... 19 41 N 97 13 E
Long Xuyen .... 10 23 N 105 25 E
Lop Buri ..... 14 48 N 100 37 E
Louang Namtha . 20 57 N 101 25 E
Louangphrabang . 19 52 N 102 08 E
Lumphăt (Lomphat) ..... 13 30 N 106 59 E
Lumut ..... 4 14 N 100 38 E
Lundu ..... 1 40 N 109 51 E
Ma-Ubin ..... 16 44 N 95 39 E
Mae Hong Son .. 19 16 N 97 56 E
Mae Sot ..... 16 43 N 98 34 E
Magwe (Magway) ..... 20 09 N 94 55 E
Maha Sarakham . 16 11 N 103 18 E
Mandalay ..... 22 00 N 96 05 E
Matu ..... 2 41 N 111 32 E
Maungdaw ..... 20 49 N 92 22 E
Mawlamyine, see Moulmein
Maymyo ..... 22 02 N 96 28 E
Meiktila ..... 20 52 N 95 52 E
Melaka (Malacca). 2 12 N 102 15 E
Mémót ..... 11 49 N 106 11 E
Mergui ..... 12 26 N 98 36 E
Mersing ..... 2 26 N 103 50 E
Minbu ..... 20 11 N 94 53 E
Minbya ..... 20 22 N 93 16 E
Minh Hoa ..... 17 47 N 106 01 E
Miri ..... 4 23 N 113 59 E
Mo Duc ..... 14 57 N 108 53 E
Mogaung ..... 25 18 N 96 56 E
Mogok ..... 22 55 N 96 30 E
Mohnyin ..... 24 47 N 96 22 E
Möng Hsan ..... 20 56 N 97 34 E
Möng Küng ..... 21 36 N 97 32 E
Monywa ..... 22 07 N 95 08 E
Moulmein (Mawlamyine) .. 16 30 N 97 38 E
Moŭng Roessei .. 12 46 N 103 27 E
Muang Khammouan (Muang Thakhek) ..... 17 24 N 104 48 E
Muang Pakxan .. 18 22 N 103 39 E
Muang Pek ..... 19 35 N 103 19 E
Muang Phôn-Hông .... 18 30 N 102 25 E
Muang Xaignabouri (Sayaboury) ... 19 15 N 101 45 E
Muang Xay ..... 20 42 N 101 59 E
Muar (Bandar Maharani) ..... 2 02 N 102 34 E
Muara ..... 5 02 N 115 04 E
Mudon ..... 16 15 N 97 44 E
Mukah ..... 2 54 N 112 06 E
Mukdahan ..... 16 32 N 104 43 E
My Tho ..... 10 21 N 106 21 E
Myanaung ..... 18 17 N 95 19 E
Myaungmya .... 16 36 N 94 56 E
Myebon ..... 20 03 N 93 22 E
Myingyan ..... 21 28 N 95 23 E
Myinmu ..... 21 56 N 95 35 E
Myitkyinā ..... 25 23 N 97 24 E
Myohaung ..... 20 35 N 93 11 E
Naba ..... 24 15 N 96 11 E
Nakhon Pathom . 13 49 N 100 03 E
Nakhon Phanom . 17 24 N 104 47 E
Nakhon Ratchasima (Khorat) ..... 14 58 N 102 07 E
Nakhon Sawan .. 15 41 N 100 07 E
Nakhon Si Thammarat .... 8 26 N 99 58 E
Nam Can ..... 8 49 N 105 01 E
Nam Dinh ..... 20 25 N 106 10 E
Namtu ..... 23 05 N 97 24 E
Nan ..... 18 47 N 100 47 E
Nang Rong ..... 14 38 N 102 48 E
Narathiwat ..... 6 26 N 101 50 E
Nghia Dan (Thai Hoa) .... 19 18 N 105 26 E
Nha Trang ..... 12 15 N 109 11 E
Ninh Binh ..... 20 15 N 105 59 E
Nong Khai ..... 17 52 N 102 44 E
Nonthaburi ..... 13 50 N 100 29 E
Nyaunglebin .... 17 57 N 96 44 E
Ŏdŏngk ..... 11 48 N 104 45 E
Pa-an (Hpa-an) .. 16 53 N 97 38 E
Pagan ..... 21 10 N 94 52 E
Pai ..... 19 19 N 98 27 E
Paillin ..... 12 51 N 102 36 E
Pak Chong ..... 14 42 N 101 25 E
Pakokku ..... 21 20 N 95 06 E
Pakxé ..... 15 07 N 105 47 E
Palaw ..... 12 58 N 98 39 E
Pathein, see Bassein
Pathum Thani ... 14 01 N 100 32 E
Pattani (Patani) .. 6 52 N 101 16 E
Pattaya ..... 12 54 N 100 51 E
Paung ..... 16 37 N 97 28 E
Paungde ..... 18 29 N 95 30 E
Péam Prus ..... 12 19 N 103 09 E
Pegu (Bago) .... 17 20 N 96 29 E
Petaling Jaya ... 3 05 N 101 39 E
Phan Rang ..... 11 34 N 108 59 E
Phan Thiet ..... 10 56 N 108 06 E
Phangnga ..... 8 28 N 98 32 E
Phatthalung .... 7 37 N 100 05 E
Phayao ..... 19 10 N 99 55 E
Phetchabun .... 16 25 N 101 08 E
Phetchaburi (Phet Buri) ..... 13 06 N 99 57 E
Phichit ..... 16 26 N 100 22 E
Phitsanulok ..... 16 50 N 100 15 E
Phnom Penh (Phnum Pénh or Pnom Penh) ... 11 33 N 104 55 E
Phnum Tbêng Méanchey .... 13 49 N 104 58 E
Phon Tiou, see Ban Phôntiou
Phôngsali ..... 21 41 N 102 06 E
Phra Nakhon Si Ayutthaya (Ayutthaya) ... 14 21 N 100 33 E
Phra Phutthabat . 14 43 N 100 48 E
Phrae ..... 18 09 N 100 08 E
Phsar Réam (Ream) ..... 10 30 N 103 37 E
Phu Cat ..... 14 01 N 109 03 E
Phu Cuong, see Thu Dau Mot
Phu Loc ..... 16 16 N 107 53 E
Phu My ..... 14 10 N 109 03 E
Phu Rieng ..... 11 40 N 106 55 E
Phu Tho ..... 21 24 N 105 13 E
Phuket ..... 7 53 N 98 24 E
Phumi Banam ... 11 19 N 105 18 E
Phumĭ Siĕmbok .. 13 17 N 105 56 E
Phuoc Long .... 9 26 N 105 28 E
Pinang, see George Town
Pleiku (Play Cu) .. 13 59 N 108 00 E
Pontian Kechil ... 1 29 N 103 23 E
Port Dickson .... 2 31 N 101 48 E
Poŭthisăt (Pursat) ..... 12 32 N 103 55 E
Prey Vêng ..... 11 29 N 105 19 E
Prome (Pye) .... 18 49 N 95 13 E
Putao ..... 27 21 N 97 24 E
Pyapon ..... 16 17 N 95 41 E
Pyawbwe ..... 20 49 N 95 44 E
Pyinmana ..... 19 44 N 96 13 E
Pyu ..... 18 29 N 96 26 E
Quang Ngai .... 15 07 N 108 48 E
Quang Tri ..... 16 45 N 107 12 E
Que Son ..... 15 40 N 108 14 E
Queenstown, *new town* ..... 1 18 N 103 48 E
Qui Chau ..... 19 33 N 105 06 E
Qui Hop ..... 19 19 N 105 09 E
Qui Nhon ..... 13 46 N 109 14 E
Rach Gia ..... 10 01 N 105 05 E
Ramree, see Kyaukpyu
Rangoon, see Yangôn
Ranong ..... 9 58 N 98 38 E
Raub ..... 3 48 N 101 52 E
Rayong ..... 12 40 N 101 17 E
Roi Et ..... 16 03 N 103 40 E
Rôviĕng Tbong .. 13 21 N 105 07 E
Sa Dec ..... 10 18 N 105 46 E
Sagaing ..... 21 52 N 95 59 E
Saigon, see Ho Chi Minh City
Sakon Nakhon ... 17 10 N 104 09 E
Salin ..... 20 35 N 94 40 E
Sam Son ..... 19 44 N 105 54 E
Samut Prakan ... 13 36 N 100 36 E
Samut Sakhon (Samut Sakorn) . 13 32 N 100 17 E
Sandakan ..... 5 50 N 118 07 E
Sândăn ..... 12 42 N 106 01 E
Sandoway ..... 18 28 N 94 22 E
Sara Buri ..... 14 32 N 100 55 E
Saravan ..... 15 43 N 106 25 E
Sarikei ..... 2 07 N 111 31 E
Sattahip ..... 12 40 N 100 54 E
Satun (Satul) .... 6 37 N 100 04 E
Savannakhét .... 16 33 N 104 45 E
Sawankhalok .... 17 19 N 99 50 E
Sayaboury, see Muang Xaignabouri
Segamat ..... 2 30 N 102 49 E
Senmonorom .... 12 27 N 107 12 E
Seremban ..... 2 43 N 101 56 E
Seria ..... 4 37 N 114 19 E
Shwebo ..... 22 34 N 95 42 E
Sibu ..... 2 18 N 111 49 E
Siĕmpang ..... 14 07 N 106 23 E
Siĕmréab ..... 13 22 N 103 51 E
Sihanoukville, see Kâmpóng Saôm
Simanggang, see Sri Aman
Singapore, *historical centre* . 1 16 N 103 50 E
Singora, see Songkhla
Sisaket (Srisaket) . 15 07 N 104 20 E
Sisŏphŏn ..... 13 35 N 102 59 E
Sittwe (Akyab) .. 20 09 N 92 54 E
Soc Trang ..... 9 36 N 105 58 E
Son La ..... 21 19 N 103 54 E
Son Tay ..... 21 08 N 105 30 E
Song ..... 2 01 N 112 33 E
Song Cau ..... 13 27 N 109 13 E
Songkhla (Singora) ..... 7 12 N 100 36 E
Sri Aman (Simanggang) .. 1 15 N 111 26 E
Srisaket, see Sisaket
Stoeng Trêng (Stung Treng) .. 13 31 N 105 58 E
Sukhothai ..... 17 01 N 99 49 E
Sungai Kolok .... 6 02 N 101 58 E
Sungai Petani ... 5 39 N 100 30 E
Suphan Buri .... 14 28 N 100 07 E
Surat Thani ..... 9 08 N 99 19 E
Surin ..... 14 53 N 103 29 E
Svay Chék ..... 13 48 N 102 58 E
Syriam ..... 16 46 N 96 15 E
Taikkyi ..... 17 19 N 95 58 E
Taiping ..... 4 51 N 100 44 E
Tak ..... 16 52 N 99 08 E
Takêv (Takéo) ... 10 59 N 104 47 E
Tam Ky ..... 15 34 N 108 29 E
Tam Quan ..... 14 35 N 109 03 E
Tamu ..... 24 13 N 94 19 E
Tan An ..... 10 32 N 106 25 E
Tăng Krâsăng ... 12 34 N 105 03 E
Tangyan ..... 22 29 N 98 24 E
Taphan Hin ..... 16 13 N 100 26 E
Tatkon ..... 20 07 N 96 13 E
Taungdwingyi ... 20 01 N 95 33 E
Taunggyi ..... 20 47 N 97 02 E
Taungup ..... 18 51 N 94 14 E
Tavoy (Dawei) ... 14 05 N 98 12 E
Tawau ..... 4 15 N 117 54 E
Tay Ninh ..... 11 18 N 106 06 E
Teluk Intan (Telok Anson) .. 4 02 N 101 01 E
Temerloh ..... 3 27 N 102 25 E
Tenasserim ..... 12 05 N 99 01 E
Tha Kham, see Ban Tha Kham
Thai Binh ..... 20 27 N 106 20 E
Thai Hoa, see Nghia Dan
Thai Nguyen .... 21 36 N 105 50 E
Thanbyuzayat ... 15 58 N 97 44 E
Thanh Hoa ..... 19 48 N 105 46 E
Thanh Tri ..... 9 26 N 105 45 E
Tharrawaddy .... 17 39 N 95 48 E
Thaton ..... 16 55 N 97 22 E
Thayetchaung ... 13 52 N 98 16 E
Thayetmyo ..... 19 19 N 95 11 E
Thoi Binh ..... 9 27 N 105 03 E
Thongwa ..... 16 46 N 96 32 E
Thu Dau Mot (Phu Cuong) ... 10 58 N 106 39 E
Thuong Duc .... 15 50 N 107 56 E
Toa Payoh, *new town* ..... 1 20 N 103 51 E
Tonzang ..... 23 36 N 93 42 E
Toungoo ..... 18 56 N 96 26 E
Tourane, see Da Nang
Trang ..... 7 33 N 99 36 E
Trat ..... 12 14 N 102 30 E
Tutong ..... 4 48 N 114 39 E
Tuy An ..... 13 17 N 109 16 E
Tuy Hoa ..... 13 05 N 109 18 E
Tuyen Hoa ..... 17 50 N 106 10 E
Tuyen Quang ... 21 49 N 105 13 E
Twante ..... 16 43 N 95 56 E
Ubon Ratchathani .... 15 14 N 104 54 E
Udon Thani ..... 17 26 N 102 46 E
Uthai Thani ..... 15 22 N 100 03 E
Uttaradit ..... 17 38 N 100 06 E
Victoria (Labuan) ..... 5 17 N 115 15 E
Vientiane (Viangchan) .... 17 58 N 102 36 E
Viet Tri ..... 21 18 N 105 26 E
Vinh ..... 18 40 N 105 40 E
Vinh Chau ..... 9 19 N 105 59 E
Virôchey ..... 13 59 N 106 49 E
Vung Tau ..... 10 21 N 107 04 E
Wakema ..... 16 36 N 95 11 E
Warin Chamrap .. 15 12 N 104 53 E
Waw ..... 17 28 N 96 41 E
Wundwin ..... 21 05 N 96 02 E
Xam Nua ..... 20 25 N 104 02 E
Xiangkhoang .... 19 20 N 103 22 E
Yala ..... 6 33 N 101 18 E
Yamethin ..... 20 26 N 96 09 E
Yandoon ..... 17 02 N 95 39 E
Yangôn (Rangoon) ..... 16 47 N 96 10 E
Yasothon ..... 15 45 N 104 08 E
Ye ..... 15 15 N 97 51 E
Ye-u ..... 22 46 N 95 26 E
Yegyi ..... 17 21 N 95 07 E
Yen Bai ..... 21 42 N 104 52 E
Yenangyaung ... 20 28 N 94 53 E
Yishun, *new town*. 1 26 N 103 51 E
Zalun ..... 17 29 N 95 34 E

Physical features and points of interest

Andaman Sea ... 10 00 N 95 00 E
Angkor, *historical site* .. 13 25 N 103 52 E
Annamese (Annamite, Annamitique or Thruong Son) Cordillera ..... 17 00 N 106 00 E
Aôral, Mount .... 12 02 N 104 10 E
Arakan Mountains ..... 19 00 N 94 40 E
Ayer Chawan Island ..... 1 16 N 103 42 E
Ayer Merbau Island ..... 1 16 N 103 43 E

Ayeyarwady, see Irrawaddy
Ba Be National Park ..... 22 23 N 105 35 E
Ba Vi National Park ..... 21 06 N 105 22 E
Bach Long Vi Island ..... 20 08 N 107 44 E
Balabac Strait ... 7 40 N 117 00 E
Bangfai, *river* .... 16 57 N 104 45 E
Banghiang, *river* . 16 03 N 105 15 E
Baram, *river* .... 4 36 N 113 59 E
Bassein, *river* ... 15 56 N 94 18 E
Batang Ai Reservoir ..... 1 05 N 111 59 E
Batu Hill ..... 2 16 N 113 43 E
Bedok Reservoir . 1 20 N 103 56 E
Belait, *river* ..... 4 35 N 114 12 E
Bengal, Bay of .. 15 00 N 90 00 E
Bia, Mount ..... 18 59 N 103 09 E
Bilauktaung Range ..... 13 00 N 99 00 E
Bintang, Mount .. 5 25 N 100 52 E
Black (Da), *river* . 21 15 N 105 20 E
Bolovens Plateau . 15 20 N 106 20 E
Borneo, *island* ... 1 00 N 114 00 E
Brunei Bay ..... 5 05 N 115 18 E
Bukit Panjang, *neighbourhood* . 1 23 N 103 46 E
Bukum Island ... 1 14 N 103 46 E
Ca (Xongka), *river* ..... 18 45 N 105 45 E
Ca Mau Point ... 8 38 N 104 44 E
Cameron Highlands ..... 4 29 N 101 23 E
Cammon Plateau . 17 50 N 105 15 E
Cardamom, see Krâvanh Mountains
Central Catchment Nature Reserve . 1 23 N 103 48 E
Central Highlands ..... 13 30 N 108 15 E
Changi, *neighbourhood* . 1 23 N 103 59 E
Changi International Airport ..... 1 22 N 103 59 E
Chao Phraya, *river* ..... 13 32 N 100 36 E
Chay, *river* ..... 21 39 N 105 12 E
Cheduba Island .. 18 48 N 93 38 E
Chi, *river* ..... 15 11 N 104 43 E
Chin Hills ..... 22 30 N 93 30 E
Chindwin (Chindwinn), *river* ..... 21 16 N 95 15 E
Chu (Xam), *river* . 19 53 N 105 45 E
City, *neighbourhood* . 1 17 N 103 51 E
Clear, see Lo
Crocker Range .. 5 40 N 116 20 E
Cuc Phuong National Park ... 20 19 N 105 38 E
Da, see Black
Dac Lac Plateau . 12 50 N 108 05 E
Dâmrei (Elephant) Mountains .... 11 00 N 104 05 E
Dangrek (Dângrêk or Dong Rak) Mountains .... 14 25 N 104 30 E
Darvel Peninsula . 4 58 N 118 30 E
Dawna Range ... 16 50 N 98 15 E
Dôn, *river* ..... 15 07 N 105 48 E
Dong Nai, *river* .. 10 45 N 106 46 E
Elephant, see Dâmrei Mountains
Endau, *river* .... 2 40 N 103 38 E
Faber, Mount ... 1 16 N 103 49 E
Fan Si Peak .... 22 18 N 103 46 E
Great Tenasserim, *river* ..... 12 24 N 98 37 E
Haihin, see Jars, Plain of
Hkakabo, Mount . 28 20 N 97 32 E
Hkok, see Kok
Hong, see Red
Huong (Perfume), *river* ..... 16 33 N 107 38 E
Inthanon, Mount . 18 35 N 98 29 E
Irrawaddy (Ayeyarwady), *river* ..... 15 50 N 95 06 E
Jars (Haihin) Plain of ..... 19 27 N 103 10 E
Johore Strait .... 1 28 N 103 48 E
Jurong, *neighbourhood* . 1 19 N 103 43 E
Kading (Theun), *river* ..... 18 19 N 104 00 E
Kaeng Krachan National Park .. 12 50 N 99 20 E
Kaladan, *river* ... 20 09 N 92 57 E
Kâmpông Saôm Bay ..... 10 50 N 103 32 E
Katong, *neighbourhood* . 1 19 N 103 54 E
Ke Ga (Varella) Point ..... 12 53 N 109 28 E
Kelantan, *river* ... 6 13 N 102 14 E
Kenyir Reservoir . 5 00 N 102 45 E
Keppel Harbour .. 1 16 N 103 50 E
Khao Laem Reservoir ..... 14 50 N 98 30 E
Khone Falls, *rapids* ..... 13 56 N 105 56 E
Khong, see Mekong
Khong, see Salween
Khorat (Korat) Plateau ..... 15 30 N 102 50 E
Khwae Noi, *river* . 14 01 N 99 32 E
Kinabalu, Mount . 6 05 N 116 33 E
Kinabalu National Park ..... 6 25 N 116 40 E
Kinabatangan, *river* ..... 5 42 N 118 23 E
Kok (Hkok), *river* . 20 15 N 100 09 E
Kong (Kŏng), *river* ..... 13 32 N 105 58 E
Korbu, Mount ... 4 41 N 101 18 E
Kra, Isthmus of .. 10 20 N 99 00 E
Kranji Reservoir .. 1 25 N 103 44 E
Kranji War Memorial ..... 1 25 N 103 45 E
Krâvanh (Cardamom) Mountains .... 12 00 N 103 15 E
Kumon Range ... 26 30 N 97 15 E
Ky Cung, *river* ... 22 20 N 106 52 E
Labuan Island ... 5 19 N 115 13 E
Labuk Bay ..... 6 10 N 117 50 E
Lam Pao Reservoir ..... 16 45 N 103 30 E
Langkawi Island ..... 6 22 N 99 48 E
Lemro, *river* ..... 20 25 N 93 20 E
Limbang, *river* ... 4 50 N 115 01 E
Linh Peak ..... 15 04 N 107 59 E
Lo (Clear or Panlong), *river* .. 21 18 N 105 25 E
Luang, Mount ... 8 31 N 99 47 E
Luang Chiang Dao, Mount .... 19 23 N 98 54 E
Luang (Sap) Lagoon ..... 7 30 N 100 15 E
Luong, Mount ... 20 40 N 104 40 E
Luong, Mount ... 21 35 N 104 17 E
Lupar, *river* ..... 1 30 N 111 00 E
Ma, *river* ..... 19 47 N 105 56 E
Mae Ping National Park ... 17 30 N 98 45 E
Makassar Strait .. 2 00 S 117 30 E
Malacca, Strait of ..... 2 30 N 101 20 E
Malay Peninsula . 6 00 N 102 00 E
Mali, *river* ..... 27 36 N 97 22 E
Mangin Range ... 24 20 N 95 42 E
Manipur Hills .... 24 30 N 94 30 E
Martaban, Gulf of ..... 16 30 N 97 00 E
Marudu Bay .... 6 45 N 116 55 E
Mekong (Khong, Mékôngk or Tien Giang), *river* ..... 10 15 N 105 55 E
Mergui Archipelago .... 12 00 N 98 00 E
Merlimau Island .. 1 17 N 103 42 E
Mokochu (Mongkrachu), Mount ..... 15 56 N 99 06 E
Mount Mulu National Park .. 4 06 N 114 59 E
Mu, *river* ..... 21 56 N 95 38 E
Muar, *river* ..... 2 03 N 102 34 E
Muda, *river* ..... 5 34 N 100 20 E
Mun, *river* ..... 15 19 N 105 30 E
Murai Reservoir .. 1 24 N 103 41 E
Murud, Mount ... 3 52 N 115 30 E
Nâga Hills ..... 26 00 N 95 00 E
Nam Bai Cat Tien National Park .. 11 13 N 107 12 E
Nan, *river* ..... 15 42 N 100 09 E
Nanyang University ..... 1 21 N 103 41 E
Negrais, Cape ... 16 02 N 94 12 E
Ngum, *river* ..... 19 49 N 101 16 E
Ngum Reservoir . 18 30 N 102 30 E
Niah Caves, *historical site* ... 3 49 N 113 47 E
Nmai, *river* ..... 25 42 N 97 30 E
Ou, *river* ..... 20 04 N 102 13 E
Pa Sak, *river* .... 14 21 N 100 35 E
Padas, *river* ..... 5 12 N 115 34 E
Pagon Peak .... 4 18 N 115 19 E
Pahang, *river* .... 3 32 N 103 28 E
Pandan Reservoir ..... 1 19 N 103 45 E
Panlong, see Lo
Pasir Panjang, *neighbourhood* . 1 18 N 103 46 E
Pâtkai (Patkoi) Range ..... 27 00 N 96 00 E
Pegu, *river* ..... 16 47 N 96 13 E
Pegu Mountains . 19 00 N 95 50 E
Penang Bridge .. 5 23 N 100 23 E
Penang (Pinang) Island ..... 5 24 N 100 14 E
Perfume, see Huong
Ping, *river* ..... 15 42 N 100 09 E
Phu Quoc Island . 10 12 N 104 00 E
Phuket Island ... 8 00 N 98 20 E
Popa, Mount .... 20 55 N 95 15 E
Poyan Reservoir . 1 23 N 103 40 E
Rajang, *river* .... 2 07 N 111 12 E
Ramree Island ... 19 06 N 93 48 E
Rangoon, see Yangôn
Rao, Mount ..... 18 09 N 105 25 E
Red (Hong), *river* . 20 17 N 106 34 E
Ron Point ..... 18 07 N 106 27 E
Sab (Sap), *river* .. 11 34 N 104 57 E
Sabah, *region* ... 5 30 N 117 00 E
Saigon (Sai Gon), *river* ..... 10 45 N 106 45 E
Sakra Island .... 1 16 N 103 42 E
Salween (Khong or Thanlwin), *river* . 16 31 N 97 37 E
Sâmkŏs, Mount .. 12 09 N 103 03 E
Sap, see Luang Lagoon
Saramati, Mount . 25 44 N 95 02 E
Sarawak, *region* . 2 30 N 113 30 E
Sarimbun Reservoir ..... 1 26 N 103 41 E
Seletar Reservoir . 1 24 N 103 48 E
Seletar River Reservoir ..... 1 25 N 103 52 E
Sembawang, *neighbourhood* . 1 27 N 103 50 E
Sentosa Island .. 1 15 N 103 50 E
Serangoon, *neighbourhood* . 1 22 N 103 54 E
Serangoon Harbour ..... 1 23 N 103 57 E
Seraya Island ... 1 16 N 103 43 E
Shan Plateau .... 22 00 N 98 00 E
Siam, see Thailand, Gulf of
Singapore, National University of ... 1 18 N 103 46 E
Singapore Island . 1 22 N 103 48 E
Singapore Strait . 1 15 N 104 00 E
Sirikit Reservoir .. 17 55 N 100 35 E
Sirinthan Reservoir ..... 15 00 N 105 55 E
Sittang (Sittoung), *river* ..... 17 10 N 96 58 E
Son Islands ..... 8 43 N 106 36 E
Songkhram, *river* ..... 17 39 N 104 28 E
South China Sea . 15 00 N 115 00 E
Srilaana National Park ..... 19 15 N 99 10 E
Srinakarin National Park ... 14 45 N 99 00 E
Srinakarin Reservoir ..... 14 45 N 99 00 E
Sudong Island ... 1 13 N 103 44 E
Sulu Sea ..... 8 00 N 117 30 E
Ta Pi, *river* ..... 9 13 N 99 24 E
Tahan, Mount ... 4 38 N 102 14 E
Taman Negara National Park ... 4 40 N 102 30 E
Tarutao National Park ..... 6 30 N 99 30 E
Tawâng, Mount .. 27 02 N 96 53 E
Tay Con, see Tsi Can, Mount
Tekong Besar Island ..... 1 24 N 104 03 E
Temengor Reservoir ..... 5 30 N 101 22 E
Tengeh Reservoir ..... 1 21 N 103 39 E
Tha, *river* ..... 20 07 N 100 36 E
Thailand (Siam), Gulf of ..... 10 00 N 102 00 E
Thanlwin, see Salween
Thap Lan National Park ... 14 10 N 102 15 E
Theun, see Kading
Three Pagodas Pass ..... 15 18 N 98 23 E
Thung Salaeng Luang National Park ..... 16 45 N 100 50 E
Tien Giang, see Mekong
Timah Hill ..... 1 21 N 103 47 E
Tioman Island ... 2 48 N 104 11 E
Tonkin, Gulf of .. 20 00 N 108 00 E
Tonle Sap (Great Lake or Tônlé Sab), *lake* ..... 13 00 N 104 00 E
Tram Chin Reserve ..... 10 40 N 105 33 E
Tsi Can (Tay Con), Mount ... 22 50 N 104 45 E
Ubin Island ..... 1 24 N 103 58 E
Ubol Ratana Reservoir ..... 16 45 N 102 30 E
Upper Peirce Reservoir ..... 1 22 N 103 48 E
Varella, Cape, see Ke Ga Point
Victoria, Mount .. 21 14 N 93 55 E
Wang, *river* ..... 17 08 N 99 02 E
Xai Lai Leng, Mount ..... 19 12 N 104 11 E
Xam, see Chu
Xiangkhoang Plateau ..... 19 30 N 103 10 E
Xongka, see Ca
Yang Sin, Mount . 12 24 N 108 26 E
Yangôn (Rangoon), *river* ..... 16 29 N 96 21 E
Yom, *river* ..... 15 52 N 100 16 E

Deforestation

Human activity has been rapidly altering the stands of virgin forest in Southeast Asia. Most deforestation results from removal for fuelwood and clearing for agriculture and grazing. Although only a relatively small portion of the total land area has been permanently cleared for cultivation—*e.g.*, in Java (Indonesia) and western Luzon (the Philippines)—in some areas shifting cultivation has brought about the replacement of virgin forest with secondary growth. In addition, nearly all countries have commercial logging industries; notable are those in Indonesia, Malaysia, Thailand, and Myanmar. A growing problem has been illegal logging. Thus, timber harvesting has come to contribute significantly to deforestation. Programs in social forestry and reforestation have yet to halt the rapid denuding of the landscape.

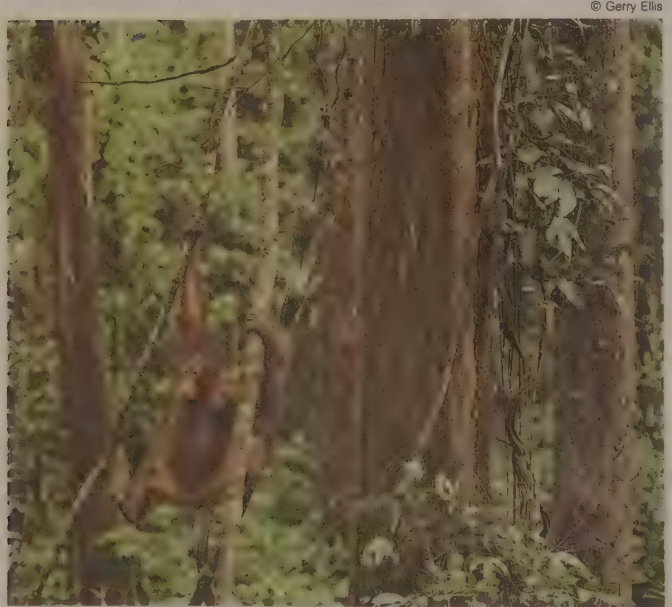
Animal life. Southeast Asia is situated where two major divisions of the world's fauna meet. The region itself constitutes the eastern half of what is called the Oriental, or Indian, zoogeographic region (part of the much larger realm of Megagaea). Bordering along the south and east is the Australian zoogeographic region, and the eastern portion of insular Southeast Asia—Celebes (Sulawesi), the Moluccas, and the Lesser Sunda Islands—constitutes a transition zone between these two faunal regions.

Southeast Asia is notable, therefore, for a considerable diversity of wildlife throughout the region. These differences are especially striking between the species of the eastern and western fringes as well as between those of the archipelagic south and the mainland north. The differences stem largely from the isolation, over varying lengths of geologic time, of species following their migration from the Asian continent. In addition, the tropical rain forests in many parts of the region, with their great diversity of vegetation, have made possible the development of complex communities of animals that fill specialized ecological niches. Especially numerous are arboreal and flying creatures.

The distinction between the two faunal regions is best depicted by their mammal populations. In general, Australia is inhabited largely by marsupials (pouched mammals) and monotremes (egg-laying mammals), while Southeast Asia contains placental mammals and such hybrid species as the bandicoot of eastern Indonesia. Small mammals such as monkeys and shrews are the most numerous, while in many areas the larger mammals have been pushed into more remote areas and national preserves. Bears, gibbons, elephants, deer, civets, and pigs are found in both mainland and insular Southeast Asia, as are diminishing numbers of tigers. The Malayan tapir, a relative of the rhinoceros, is native to the Malay Peninsula and Sumatra, while the tarsier is found in the Philippines and parts of Indonesia. A number of rare endemic species are found in Indonesia and East (insular) Malaysia, including the Sumatran and Javan rhinoceros, the orangutan, the anoa (a dwarf buffalo), the babirusa (a wild swine), and the palm civet.

Rare species

As the pace of development accelerates and populations continue to expand in Southeast Asia, concern has increased regarding the impact of human activity on the region's environment. A significant portion of Southeast Asia, however, has not changed greatly and remains an unaltered home to wildlife. The nations of the region, with only few exceptions, have become aware of the need to maintain forest cover not only to prevent soil erosion but to preserve the diversity of flora and fauna. Indonesia, for example, has created an extensive system of national parks and preserves for this purpose. Even so, such species as the Javan rhinoceros face extinction, with only a handful of the animals remaining in western Java.



A young orangutan in the tropical-evergreen forest of northern Borneo, Sabah, East Malaysia.

THE PEOPLE

By the late 20th century, Southeast Asia's population (including Indonesia and the Philippines) was approaching a half billion, or about one-twelfth of the world's total. This population, however, was unevenly distributed within the region. By far the nation with the largest population was Indonesia, with about two-fifths of the regional total; in contrast, Brunei's population was only a tiny fraction of that. Nearly half of the regional population was accounted for by the mainland states, with Vietnam and Thailand being the most populous.

Settlement patterns. Southeast Asia is predominantly rural: three-fourths of the people live in nonurban areas. Moreover, population is heavily clustered in fertile river valleys and especially in delta areas, such as those of the Mekong and Irrawaddy rivers. Historical, cultural, and environmental influences also have affected the settlement patterns. Java and other core areas such as the Bangkok (Thailand), Hanoi, and Manila metropolitan areas contain high population densities.

While the rate of urbanization in Southeast Asia is relatively low compared with those of other developing regions, it is increasing rapidly. Singapore is unique in that it is essentially totally urban. In addition, the Philippines has a much higher than average level of urbanization, in part because of its Spanish and American colonial history. The largest cities—Jakarta (Indonesia), Bangkok, and Manila—are among the world's most populous. The growth of cities of all sizes is being fueled primarily by natural increase, but rural-urban migration also is a significant contributor. Rural dwellers continue to be attracted by the promise of employment and other opportunities, but for many migrants the informal (undocumented) economic sector in these large cities is the only hope for some form of employment.

Settlement patterns in rural areas tend to be associated with agricultural practices. Shifting cultivation is still common in some parts of the region (notably the remote interior areas of Myanmar, Vietnam, and the island of Borneo), although the amount of land so utilized is gradually shrinking. The village is the unit of settlement and often functions collectively, and typically it is moved from time to time. By contrast, wet-rice cultivation, the dominant form of agriculture in Southeast Asia, is sedentary and results in relatively large rural agglomerations with well-developed village life and customs. Dry and upland farming often produces scattered homesteads.

Resettlement

Population resettlement to provide agricultural employment and access to land is important in some Southeast Asian countries, notably Indonesia, Malaysia, and Vietnam. By far the largest program has been conducted in Indonesia, where more than four million people have been voluntarily resettled from Java and Bali to the less populated islands. Despite considerable success, the program has been plagued by such problems as improper site selection, environmental deterioration, migrant adjustment, land conflicts, and inadequate financing. A program in Malaysia also has been quite successful, in part because it has set much smaller resettlement targets and has been

better funded. Vietnamese development policy also has utilized the resettlement of people in an effort to revitalize areas outside the major population centres.

Ethnic composition. Southeast Asia's population includes a wide variety of ethnic groups and cultures. This diversity is related to its position as a focus of converging land and sea routes. In addition, over the span of human habitation, the region alternately has been a bridge and a barrier to the movement of people. The peopling of Southeast Asia took place through various southward migrations. The initial peoples arrived from the Asian continental interior. Successive movement displaced these initial settlers and created a complex ethnic pattern.

On the mainland the Khmer peoples of Cambodia remain as ancestors of earlier Pareoean peoples. Similarly, remnants of the Mon group are found in parts of Myanmar and Thailand; the ethnic mixture there has been produced by overlaying Tibeto-Burman and Tai, Lao, and Shan peoples. The contemporary Vietnamese population originated from the Red River area in the north and may be a mixture of Tai and Malay peoples. Added to these major ethnic groups are such less numerous peoples as the Karens, Chins, and Nāgas in Myanmar, who have affinities with other Asiatic peoples. Insular Southeast Asia contains a mixture of descendants of Proto-Malay (Nesiot) and Pareoean peoples who were influenced by Malayo-Polynesian and other groups. In addition, Arabic, Indian, and Chinese influences have affected the ethnic pattern of the islands.

In modern times the Burmans account for more than two-thirds of the ethnic stock of Myanmar, while ethnic Thais and Vietnamese account for about four-fifths of the respective populations of those countries. Indonesia is clearly dominated by the Javanese and Sundanese ethnic groups, while Malaysia is more evenly split between the Malays and the Chinese. Within the Philippines, the Tagalog, Cebuano, Ilocano, and Bicol groups are significant.

Four major language families

Linguistic composition. Language patterns in Southeast Asia are highly complex and are rooted in four major language families: the Sino-Tibetan, Tai, Austro-Asiatic, and Austronesian (Malayo-Polynesian). Languages derived from the Sino-Tibetan group are found largely in Myanmar, while forms of the Tai group are spoken in Thailand and Laos. Austro-Asiatic languages are spoken in Cambodia, Laos, and Vietnam. The languages of Malaysia, Indonesia, and the Philippines are rooted in an Austronesian and Polynesian stock. Despite this broad generalization, it must be noted that innumerable separate languages as well as dialects are used in the region. This linguistic diversity is especially conspicuous in fragmented areas such as the Philippines and Indonesia and in highland and remote areas on the mainland, and it has been a retarding factor in national integration and development. Notable in this regard is Myanmar.

Dominant languages do exist in most of the nations. Burmese and Thai are spoken by large groups of people in Myanmar and Thailand, respectively. Similarly, Khmer is the primary language in Cambodia, as is Vietnamese in Vietnam. Within the Philippines, Pilipino (Filipino) and English are the official languages, but Tagalog and Visayan also are important. Malay and Indonesian are, respectively, the official languages of Malaysia and Indonesia; these languages are quite similar and are mutually intelligible. Indonesian is a good example of a true national language and is spoken widely across the archipelago. Thus, unlike in Myanmar, language actually has been a unifying element in the country.

Numerous languages also have been introduced into the region by immigrant populations. Perhaps most significant are the variety of dialects spoken by the Chinese communities in many Southeast Asian countries. The most commonly used are Cantonese, Hokkien, Hakka, and Teochew, reflecting the southern Chinese coastal origins of many of the immigrants. The largest concentration of Chinese speakers is in Singapore, where they constitute the majority population. Concentrations of ethnic Chinese also live in most of the larger urban areas of the region.

Indian immigrants also are numerous and are associated with the economic development of several Southeast Asian nations. Their role as labourers on the rubber plantations of Malaysia is well known, and Tamil and Hindi speakers form significant minorities in the country. Indian communities also are scattered throughout the region and are especially conspicuous in Singapore and Myanmar.

Religions. Buddhism, Islām, and Christianity are all practiced within Southeast Asia. Buddhism, particularly the more orthodox Theravāda form, dominates the religious pattern of most of the mainland; only in northern Vietnam is the more liberal Mahāyāna Buddhism more common.

Islām

Islām is predominant in the southern half of the Malay Peninsula, the Malay Archipelago, and the southern Philippines. As a result of the large Muslim population in Indonesia, Islām is the religion of some two-fifths of Southeast Asians. The diffusion of the religion began in the early 14th century through contact with Muslim traders in northern Sumatra. Perhaps more than any of the other religions, Islām has been a strong force in binding together its adherents. It has profoundly affected cultural, social, political, and economic matters in areas where it is practiced.

The spread of Christianity came with European contact. Roman Catholicism was introduced to insular Southeast Asia by the Spanish and the Portuguese in the 16th century and somewhat later to the Indochinese Peninsula by the French. Catholicism is most important in the Philippines and southern Vietnam. Protestantism also is locally important. The Batak and Minangkabau peoples in Sumatra and a growing number of Chinese in Singapore and elsewhere adhere to various Protestant denominations.

Hinduism, once much more widespread, now is practiced by many people in the region's Indian communities. In addition, this religion, modified by animism and other influences, is the primary faith on the island of Bali in Indonesia. Various forms of animism also are practiced in the region's more remote areas, particularly in central Borneo, northern Laos, and northern Myanmar.

Demographic trends. The annual rate of natural increase in Southeast Asia averages slightly higher than the annual world rate. Considerable variation exists, however, among the region's countries. The Philippines, Laos, Malaysia, Vietnam, and Brunei are characterized by higher growth; Singapore, Thailand, and Indonesia, on the other hand, have considerably lower rates, primarily because of the implementation of effective family-planning programs in these countries. In general, the pace of fertility decline is accelerating, although it is being offset by declining infant mortality and increasing life expectancy. Infant mortality for the region approximates the world average. In the more developed nations—especially Singapore, Malaysia, and Thailand—health care programs for infants and children have helped bring about mortality rates well below world averages, while the scarcity of these programs in such countries as Cambodia and Laos has contributed to continued high rates. Life expectancy in the region is somewhat below the world average, with Cambodia having the lowest average and Singapore the highest.

Population change also is directly related to internal and external migration. As noted above, rural-to-urban migration continues to be a major aspect of change in nearly all Southeast Asian nations. In certain countries, considerable evidence exists for movements between rural areas (*e.g.*, Thailand) and mobility between urban areas (Indonesia). Internal migration in the Philippines is dominated by movements to Manila and to the frontier areas in the south. Perhaps most significant, given the increasing mobility of the population and access to transport services, is the growth of nonpermanent population movements. Seasonal and other forms of circular migration for limited periods of time are conspicuous, especially in Malaysia, Indonesia, and Thailand. The growth in transport access also has created greater commuting ranges for individuals who in the past often had to leave their homes and fields for extended periods to take up work.

Refugee movements

Refugee movements have been conspicuous in the region, particularly since the mid-1970s. The Vietnamese out-migration to Malaysia, Thailand, Hong Kong, and Indonesia is noteworthy. Cambodian and Laotian peoples

also have experienced displacement. In addition, there have been numerous instances of religious minorities fleeing persecution, such as the departure of Muslim Burmans in the early 1990s.

THE ECONOMY

Even prior to the penetration of European interests, Southeast Asia was a critical part of the world trading system. A wide range of commodities originated in the region, but especially important were such spices as pepper, ginger, cloves, and nutmeg. The spice trade initially was developed by Indian and Arab merchants, but it also brought Europeans to the region. First the Portuguese, then the Dutch, and finally the British and French became involved in this enterprise in various countries. The penetration of European commercial interests gradually evolved into annexation of territories, as traders lobbied for an extension of control to protect and expand their activities. As a result, the Dutch moved into Indonesia, the British into Malaya, and the French into Indochina.

Europe's interest and activity in the region was further enhanced by the opening of the Suez Canal, the development of telegraphic communications, the adoption of steam shipping, and the prospects for trade with China. In the case of Malaya, the gradual diffusion of British administration provided systems of law and order and of taxation and allowed for the gradual development of infrastructure, principally reliable transport systems. This environment attracted Chinese immigrants, and the growth of the tin mining industry soon followed. Later rubber plantations were established, which brought about still further immigration. Similar developments took place in Burma (Myanmar), Vietnam, and Indonesia. In Siam (Thailand) during the second half of the 19th century, a rapid expansion of Western enterprise occurred, though not by colonization. Both British and American firms began trading in the region. The impact of the Western activity was essentially to remove trade from what had been a Chinese monopoly and to emphasize the export of a single commodity, rice. Established indigenous textile and sugar-processing industries were replaced by imports, and the economy slowly became dependent on rice exports. The Philippines gradually developed a plantation farming system under Spanish and later American influence, although rice, sugar, and tobacco continued to be produced by small-scale growers and processed by Chinese enterprises until the mid-19th century.

The incorporation of Southeast Asia into the world economy had a major impact on the distribution of the region's economic development, and it created more uneven patterns of population growth and economic activity. It also brought about a stronger sense of class distinction and resulted in a larger discrepancy between the wealthy and poor. The worldwide economic depression of the 1930s severely affected the commercialized areas most dependent on the world economy. Unemployment rose, and the period produced the seeds of political change and activism that culminated in the independence of most of the region's countries after World War II. Effects of the depression

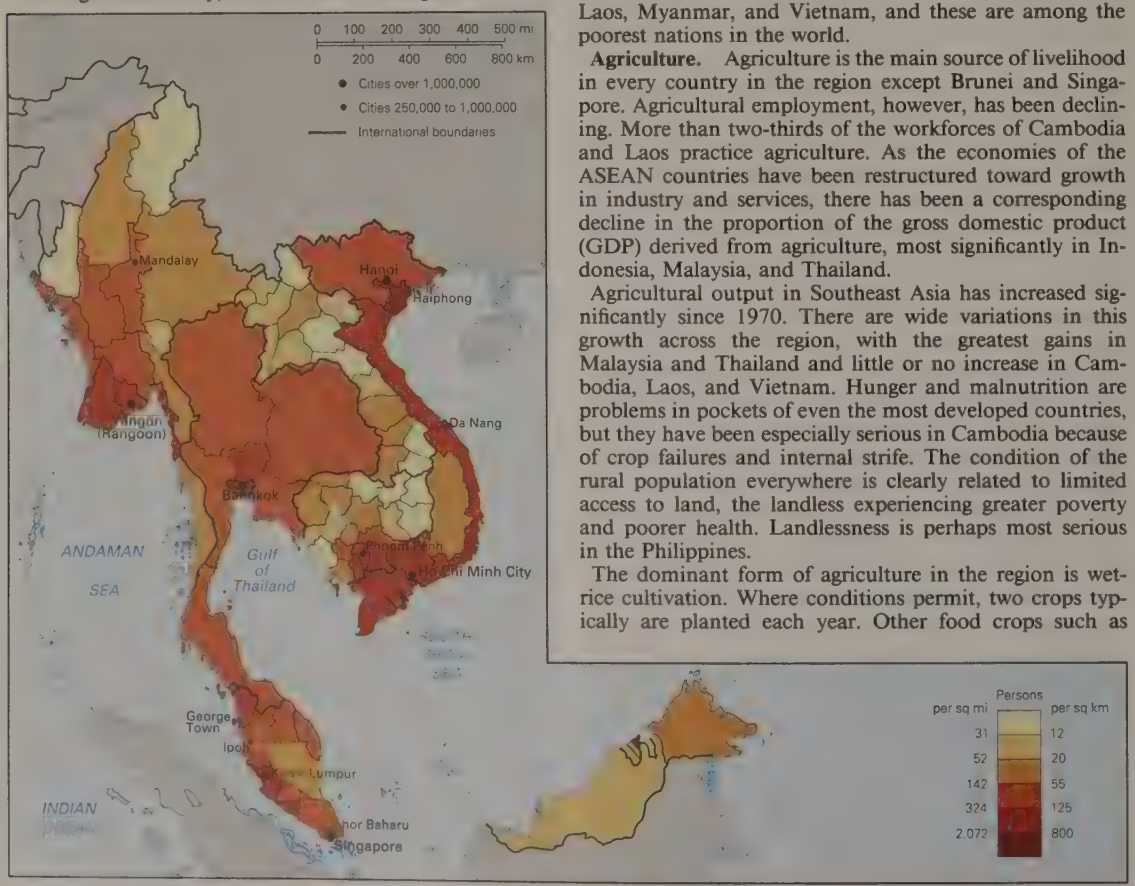
Since the 1950s the economic development strategies of virtually all the capitalist Southeast Asian states have emphasized urban industrialization, while agricultural development generally has been viewed as subsidiary to industrial growth. These strategies have met with mixed success. Indeed, the trading pattern of the region by and large has continued to be one of producing and exporting raw materials and importing manufactured goods. Only Singapore has reached an advanced level of industrialization, in the process becoming one of the world's great centres of industry and commerce.

There is great disparity in development rates within the region, especially between the member and nonmember countries of the Association of Southeast Asian Nations (ASEAN). Those belonging to this grouping—Brunei, Indonesia, Malaysia, the Philippines, Singapore, and Thailand—generally have experienced significant economic development since the mid-1960s; the exception has been the Philippines, the economy of which has grown at a much slower rate. Development has been extremely slow or nonexistent in the non-ASEAN countries of Cambodia, Laos, Myanmar, and Vietnam, and these are among the poorest nations in the world.

Agriculture. Agriculture is the main source of livelihood in every country in the region except Brunei and Singapore. Agricultural employment, however, has been declining. More than two-thirds of the workforces of Cambodia and Laos practice agriculture. As the economies of the ASEAN countries have been restructured toward growth in industry and services, there has been a corresponding decline in the proportion of the gross domestic product (GDP) derived from agriculture, most significantly in Indonesia, Malaysia, and Thailand.

Agricultural output in Southeast Asia has increased significantly since 1970. There are wide variations in this growth across the region, with the greatest gains in Malaysia and Thailand and little or no increase in Cambodia, Laos, and Vietnam. Hunger and malnutrition are problems in pockets of even the most developed countries, but they have been especially serious in Cambodia because of crop failures and internal strife. The condition of the rural population everywhere is clearly related to limited access to land, the landless experiencing greater poverty and poorer health. Landlessness is perhaps most serious in the Philippines.

The dominant form of agriculture in the region is wet-rice cultivation. Where conditions permit, two crops typically are planted each year. Other food crops such as Rice



Population density of Southeast Asia.

corn (maize), cassava, and pulses (legumes) frequently are grown in drier areas where there is too little water for a second planting of rice. Rice production requires a reliable water supply. Thailand and the Philippines rely heavily on rain-fed systems, while Indonesia utilizes irrigation to a large extent. Irrigation or some other form of water control is especially critical in the cultivation of the high-yielding varieties (HYVs) of rice that have been introduced since the 1960s. The spread of the so-called Green Revolution—in which HYVs and chemical fertilizers and pesticides are utilized—has brought mixed results. There is little doubt that production has increased because of the higher yields of these hybrid strains and because their more rapid maturation increases the possibility of multiple annual crops. Frequently, however, poorer farmers are not able to take advantage of these strains, because of the high cost of their use. The goal of rice self-sufficiency has been difficult to achieve for most countries.

A large variety of cash crops are grown for the local and export markets, both on large commercial estates and by individual growers or smallholders. Tree crops are the most important in terms of value, although the area devoted to them is limited largely to equatorial areas. Rubber and palm oil are significant in Malaysia, Indonesia, and southern Thailand, while coconuts and sugar are important in the Philippines. Other major export crops are cacao, coffee, and spices, while crops grown largely for local and regional consumption include chilies, sweet potatoes, peanuts (groundnuts), and tobacco. The cultivation of opium poppies is important in parts of Myanmar and Thailand.

The emphasis on rubber and palm oil production is in response to a considerable (though fluctuating) worldwide demand for these commodities and because of a nearly continuous harvest period that provides year-round employment. Foreign corporations once dominated production, but, as the region's countries gained independence, much of the production was nationalized. Government ownership continues to predominate, with increasing private ownership.

Fishing contributes only a token amount to the GDP of Southeast Asian countries, but it is an important livelihood in certain areas and supplies a significant portion of the local diet. Marine output has gradually expanded with new technologies. The maritime nations of Thailand, Indonesia, Malaysia, and the Philippines all have globally important fishing industries. Shrimp catches are especially in demand in the world economy. Aquaculture has become increasingly important in the region, such species as shrimp, carp, and grouper being raised in excavated ponds.

Industry. Industrialization in Southeast Asia is a relatively recent phenomenon, much of the development having occurred only since the early 1960s. As mentioned above, industrialization policies have been critical goals in the market economies of the ASEAN countries; and, in all of them except Brunei, industry's share of the GDP has grown considerably. The most significant increases have occurred in Singapore, Thailand, and the Philippines. Manufacturing in particular has accounted for the greatest changes, with Indonesia, Malaysia, and Thailand making especially large gains during the 1980s.

Growth in manufacturing

Small factories dominate, both in terms of the number of companies and the number of workers employed. Agricultural processing is most important in virtually all nations. The notable exception is Singapore, where the manufacture of a variety of products, headed by electrical and electronic and transport equipment, is dominant. In Thailand, Myanmar, and the Philippines, textiles and clothing are significant, as is the chemical industry in Thailand and Indonesia. Light, labour-intensive goods, such as electrical and electronic products, are increasingly important. It is in the manufacture of these products and textiles that the most employment has been gained.

Tin is the most important metallic mineral in the region in terms of value, and Thailand, Malaysia, and Indonesia account for more than half of world production. In Malaysia and elsewhere, however, alluvial lodes are becoming depleted, and the remaining concentrations are less economical to mine. Fluctuating market prices have also discouraged tin production. Nickel, copper, and chromite are also mined, although the quantities produced in the region are minor in terms of world production. Southeast Asia has considerable reserves of oil and natural gas, notably in Indonesia, Malaysia, and Brunei.

Trade. Given Southeast Asia's strategic location and the early development of trade there, it is not surprising that trade is especially important to all nations in the region. The value of regional trade is about one-third that of the United States. Most striking is the almost total dominance of trade by the market economies. Exports, as a percentage of the GDP, are small in Cambodia, Myanmar, Vietnam, and Laos and moderately so in Thailand, the Philippines, and Indonesia. Countries with a relatively large proportion of export trade are Singapore, Malaysia, and Brunei. Composition of exports is important. In this respect, Indonesia—the trade structure of which long has been dominated by oil—has been relatively successful in diversifying its exports toward plywood, rattan, coffee, rubber, and textiles. Conversely, Malaysia, with a trade pattern of exporting palm oil, tropical hardwoods, and tin, now derives the majority of its export income from petroleum products. This revenue has been used to build up the country's industrial base. Thailand exhibits a much less diverse export structure, where food and manufactured goods account for nearly all of its total trade. Likewise, Brunei relies almost entirely on its petroleum exports. Singapore, however, has utilized its unique geographic position and highly educated labour force to attract multinational corporations. As a result, investment in the manufacturing and, increasingly, service sectors has greatly expanded.

Intraregional trade among the ASEAN members, while important, accounts for only about one-fifth of Southeast Asia's total trade. Philippine trade within the region is especially small, reflecting its long-term orientation toward the United States. Far more important, therefore, is the trade with countries outside the region, dominated by that with Japan, Europe, and the United States; increasingly significant, however, is the trade with Taiwan, Hong Kong, and South Korea.

Trading partners

Transportation and communications. Before World War II the various colonial powers of the region attempted to provide reliable transport systems. Emphasis first was placed on developing road networks, followed by railways. The infrastructure that was built during the colonial period, however, deteriorated rapidly after the war; since achieving independence, many of the countries gradually have been restoring and extending their road networks. This activity has been notable in Indonesia, where, because of the country's vastness, the task has been enormous. Transport systems in Myanmar and the countries of the Indochinese Peninsula in general are poorly developed, except in some parts of Vietnam, where improvements were made during wartime.

Road transport continues to be of overwhelming importance in the region. Since all countries but Laos have maritime access, water transport is next in importance. It is especially vital in archipelagic Indonesia and the Philippines and also is significant in Malaysia and Thailand. Railways are of minor importance, in part because the region's archipelagic nature is not conducive to their construction but more critically because the relatively short hauling distances allow road transport to be more competitive. Even in Thailand—where the potential for rail transport is greatest—an extensive highway system and the availability of reliable vehicles provide a formidable challenge to rail.

All of the ASEAN countries have strong domestic air transport systems. The most extensive is in Indonesia, which provides critical links between the islands. In addition, the Indonesian government maintains subsidized air services to the smaller islands. Most ASEAN nations also have international air fleets, the largest of which are maintained by Singapore, Malaysia, and Thailand.

There has been increased emphasis on the development of communications throughout the ASEAN states. Singapore has become renowned for its extensive communications infrastructure and capability. Telephone service is

most abundant in the urban areas of the more developed states, although telecommunications in the rural areas of the Philippines, Indonesia, and Thailand remains deficient. Indonesia has made significant improvements in its communications infrastructure through the deployment of satellites that enhance television and telephone transmission to remote areas of the archipelago. (T.R.L.)

History

EARLY SOCIETY AND ACCOMPLISHMENTS

Origins. Knowledge of the early prehistory of Southeast Asia has undergone exceptionally rapid change as a result of archaeological discoveries made since the 1960s, although the interpretation of these findings has remained the subject of extensive debate. Nevertheless, it seems clear that the region has been inhabited from the earliest times. Hominid fossil remains date from approximately 1,500,000 years ago, and those of *Homo sapiens* from approximately 40,000 years ago. Furthermore, until about 7000 BC the seas were some 150 feet (50 metres) lower than they are now, and the area west of Makassar Strait consisted of a web of watered plains that sometimes is called Sundaland. These land connections perhaps account for the coherence of early human development observed in the Hoabinhian age, which lasted from about 13,000 to 5000 or 4000 BC. The stone tools used by hunting and gathering societies across Southeast Asia during this period show a remarkable degree of similarity in design and development. When the sea level rose to approximately its present level about 6000 BC, conditions were created for a more variegated environment and, therefore, for more extensive differentiation in human development. While migration from outside the region may have taken place, it did not do so in a massive or clearly punctuated fashion; local evolutionary processes and the circulation of peoples were far more powerful forces in shaping the region's cultural landscape.

Technological developments and population expansion. Perhaps because of a particular combination of geophysical and climatic factors, early Southeast Asia did not develop uniformly in the direction of increasingly complex societies. Not only have significant hunting and gathering populations continued to exist into the 20th century, but the familiar cultural sequences triggered by such events as the discovery of agriculture or metallurgy do not seem to apply. This is not to say that the technological capabilities of early Southeast Asian peoples were negligible, for sophisticated metalworking (bronze) and agriculture (rice) were being practiced by the end of the third millennium BC in northeastern Thailand and northern Vietnam, and sailing vessels of advanced design and sophisticated navigational skills were spread over a wider area by the same time or earlier. Significantly, these technologies do not appear to have been borrowed from elsewhere but were indigenous and distinctive in character.

Rate of societal change

These technological changes may partially account for two crucial developments in Southeast Asia's later prehistory. The first is the extraordinary seaborne expansion of speakers of Proto-Austronesian languages and their descendants, speakers of Austronesian (or Malayo-Polynesian) languages, which occurred over a period of 5,000 years or more and came to encompass a vast area and to stretch nearly half the circumference of the Earth at the Equator. This outward movement of people and culture was evolutionary rather than revolutionary, the result of societal preference for small groups and a tendency of groups to hive off once a certain population size had been reached. It began as early as 4000 BC, when Taiwan was populated from the Asian mainland, and subsequently it continued southward through the northern Philippines (3rd millennium BC), central Indonesia (2nd millennium BC), and western and eastern Indonesia (2nd and 1st millennia BC). From approximately 1000 BC on the expansion continued both eastward into the Pacific—where that immense region was populated in a process continuing to about AD 1000 as voyagers reached the Hawaiian Islands and New Zealand—and westward—where Malay peoples reached and settled the island of Madagascar sometime between AD 500 and 700, bringing with them (among other things) bananas, which are native to Southeast Asia. Thus, for a considerable period of time, the Southeast Asian region contributed to world cultural history, rather than merely accepting outside influences, as frequently has been suggested.

The second development, which began possibly as early as 1000 BC, centred on the production of fine bronze and the fashioning of bronze-and-iron objects, particularly as they have been found at the site in northern Vietnam known as Dong Son. The earliest objects consisted of socketed plowshares and axes, shaft-hole sickles, spearheads, and such small items as fishhooks and personal ornaments. By about 500 BC the Dong Son culture began producing the bronze drums for which it is known. The drums are large objects (some weigh more than 150 pounds [70 kilograms]), and they were produced by the difficult lost-wax casting process and decorated with fine geometric shapes and depictions of animals and humans. This metal industry was not derived from similar industries in China or India. Rather, the Dong Son period offers one of the most powerful—though not necessarily the only or earliest—examples of Southeast Asian societies transforming themselves into more densely populated, hierarchical, and centralized communities. Since typical drums, either originals or local renditions, have been found throughout Southeast Asia and since they are associated with a rich trade in exotics and other goods, the Dong Son culture also suggests that the region as a whole consisted not of isolated, primitive niches of human settlement but of a variety of societies and cultures tied together by broad and long-extant trading patterns. Although none of these societies possessed writing, some displayed considerable sophistication and technological skill; and, although none appears to have constituted a territorial, centralized state, new and more complex polities were forming.

Dong Son culture

Influence of China and India. Between approximately 150 BC and AD 150, most of Southeast Asia was first influenced by the more mature cultures of its neighbours to the north and west. Thus began a process that lasted for the better part of a millennium and fundamentally changed Southeast Asia. In some ways the circumstances were very different. China, concerned about increasingly powerful chiefdoms in Vietnam disturbing its trade, encroached into the region and by the end of the 1st century BC had incorporated it as a remote province of the Han empire. For generations, the Vietnamese opposed Chinese rule, but they were unable to gain their independence until AD 939. From India, however, there is no evidence of conquests, colonization, or even extensive migration. Indians came to Southeast Asia, but they did not come to rule; and no Indian power appears to have pursued an interest in controlling a Southeast Asian power from afar, a factor that may help to explain why only the Vietnamese accepted the Chinese model. Yet, in other ways the processes of Indianization and Sinicization were remarkably similar. Southeast Asia already was socially and culturally diverse, making accommodation easy. Furthermore, indigenous peoples shaped the adaption and adoption of outside influences and, indeed, seem to have sought out concepts and practices that enhanced rather than redirected changes already underway in their own societies. They also rejected some components: for example, some of the vocabulary and general theories related to the Indian notions of social hierarchy were borrowed but much of the specific practices were not, and neither Indian nor Chinese views of women as socially and legally inferior were accepted. In the later stages of the assimilation process—particularly in the Indianized areas—local syncretism often produced exuberant variations, which, despite familiar appearances, were expressions of local genius rather than just inspired borrowings.

Still, Chinese and Indian influences were anything but superficial. They provided writing systems and literature, systems of statecraft, and concepts of social hierarchy and religious belief, all of which were both of intrinsic interest and pragmatic significance to Southeast Asians of the day. For elites seeking to gain and retain control over larger and more complex populations, the applications of these

Artistic expression

ideas were obvious; but it would also seem that the sheer beauty and symbolic power of Hindu and Buddhist arts tapped a responsive vein in the Southeast Asian soul. The result was an imposing array of architectural and other cultural wonders, at first very much in the Indian image and hewing close to current styles and later in more original, indigenous interpretations. The seriousness and profundity with which all this activity was undertaken is unmistakable. By the 7th century AD Palembang in southern Sumatra was being visited by Chinese and other Buddhist devotees from throughout Asia, who came to study doctrine and to copy manuscripts in institutions that rivaled in importance those in India itself. Later, beginning in the 8th century, temple and court complexes of surpassing grandeur and beauty were constructed in central Java, Myanmar, and Cambodia; the Borobudur of the Śailendra dynasty in Java, the myriad temples of the Burman dynastic capital of Pagan, and the monuments constructed at Angkor during the Khmer empire in Cambodia rank without question among the glories of the ancient world.

Rise of indigenous states. In the realm of politics, Indian influence accompanied the rise of new political entities, which, since they do not readily fall under the Western rubric of "states," have been called *mandalas*. The *mandala* was not so much a territorial unit as a fluid field of power that emanated, in concentric circles, from a central court and depended for its continued authority largely on the court's ability to balance alliances and to influence the flow of trade and human resources. Such a conception of political organization already had surfaced among Southeast Asians, but Indian civilization provided powerful metaphors for the change underway and for ways of extending it. The *mandala* was the predominant form of the Southeast Asian state until it was displaced in the 19th century.

Between approximately the 2nd century BC and the 6th century AD, *mandala* polities appeared throughout Southeast Asia in the major river valleys and at strategic landfalls for sea traffic—generally, locations where routes for local and international trade crossed. These communities took different forms, depending on their physical setting. For example, walled and moated settlements predominated in much of the mainland but do not seem to have been constructed in insular Southeast Asia. Yet they served similar purposes to and frequently shared characteristics with *mandalas* in the same immediate region. *Mandala* sites have been located in the Mekong, Chao Phraya, and Irrawaddy river valleys; along the coasts of central Vietnam, western and northern Java, and eastern Borneo; and on the Isthmus of Kra. One of the most intriguing sites,

Oc Eo

called Oc Eo, is in the Mekong delta region of southern Vietnam. This port settlement, which flourished between the 1st and 6th centuries AD amid a complex of other settlements connected by canals (some up to 60 miles long), was not only an extraordinarily rich emporium dealing in articles from as far as Rome and inner Asia, but it was also a local manufacturing centre producing its own jewelry, pottery, and other trade goods. Almost certainly it also fed itself from wet-rice agriculture practiced in the surrounding delta. Little is known, however, about the nature of state structure in Oc Eo, although it seems to have been one of—and perhaps was prime among—an assemblage of local *mandala*-type principalities.

After the 6th century there emerged a number of larger and more powerful *mandala* states, principally in Cambodia, Myanmar, Sumatra, and Java. Often designated kingdoms or empires, these states nevertheless functioned and were structured upon the same principles that had governed their predecessors. They were, in some respects, unstable and prone to fluctuation because of shifting relations with outside powers and constant internal struggles for the position of overlordship, but they also were remarkably durable. No two states were exactly alike, each occupying a particular ecological niche and exploiting a particular combination of opportunities to survive by trade, agriculture, and war. The cultural impact of their courts long outlasted their political grasp and continued to inform their societies until modern times. Perhaps the outstanding example of this durability is Śrīvijaya, the great Sumatran trading "empire" that dominated much of Southeast Asian commerce from about the 7th to the 13th century. Śrīvijaya does not appear to have been heavily urbanized or to have had a continuously occupied capital during its roughly 700 years of existence, nor does it seem to have possessed boundaries and clearly delineated territories. Its armies, while they could be mustered and quickly dispatched overseas, were weapons of limited use. Instead, Śrīvijaya maintained its authority in a shifting and extremely varied trading world largely by means of a shrewd brand of cultural and economic politics that involved, among other things, offering a protective and mutually beneficial trading environment to all comers and maintaining a courtly culture from which the idiom of overlordship issued grandly and convincingly. Śrīvijaya was ruled by a formula supple enough to attract trade from all quarters and to exploit it at the same time.

The Khmer state

Whatever the achievements of Śrīvijaya, the Khmer (Cambodian) state that flourished in the Tonle Sap region roughly between the 9th and mid-13th centuries is widely regarded as the most impressive of the concentrically arranged ancient Southeast Asian states. This admiration largely stems from the state's extensive architectural remains, including the renowned Angkor Thom and Angkor Wat temple complexes. In many respects, however, the Angkorian imperial achievement was singular. Though informed by the *mandala* paradigm, the Khmer carried it further and shaped it more distinctively than other Southeast Asians before or since. Not only was the *devarāja* ("a god who is king") cult a Khmer innovation, but the Khmer also uniquely transformed their physical environment to support their state. They created what may be the ancient world's most intricate and capacious system of water catchment and dispersal, making it possible to grow three or even four crops of rice annually in an area not especially hospitable to raising even one. At its zenith, Angkor may have supported a population of one million in a relatively small area, with an elite apparatus and a population of bondsmen far greater than any of Cambodia's neighbours. In achieving this, however, the Khmer state surrendered the flexibility and balance critical to the *mandala* pattern and eventually fell victim to its own brittleness. Other concentric states in early Southeast Asia rose and fell; the Khmer proved unable to revive theirs once it had fallen.

THE CLASSICAL PERIOD

Components of a new age. By about 1300 much of Southeast Asia had entered a period of transition from ancient times. No single factor can account for the disruption, which lasted longer in some places than in others. The Mongol attacks of the second half of the 13th century and the disintegration of Khmer and Śrīvijayan power undoubtedly were of significance, but less dramatic changes, such as slowly changing trade patterns and political competition, may also have played an important role. Whatever the case, the shifts were not of a type or severity to bring about major disruptions; they instead paved the way for the coalescing of what can best be termed a classical age. In this period the major civilizations of Southeast Asia achieved a broader influence and greater coherence than before. They integrated rival political and cultural forms into their own, and the patterns they established were widely imitated by smaller powers that were drawn into their orbit. Regional and international trade reached a high level of development, bringing greater well-being to larger numbers of Southeast Asians than ever before. It also was an age of great change and challenges—especially in the form of new and often foreign religious, political, and economic influences—and one of constant warfare. But it was a measure of the confidence and balance of the era that these influences were absorbed and digested with little difficulty, leaving more than a millennium of creative synthesis essentially undisturbed until as late as the end of the 18th century. Many Southeast Asian civilizations can be said to have reached their definitive premodern shape during this "golden" age, which also is modern scholarship's best source of information on the classical cultures

of the region before the ravages of 19th- and 20th-century colonialism.

Five major powers

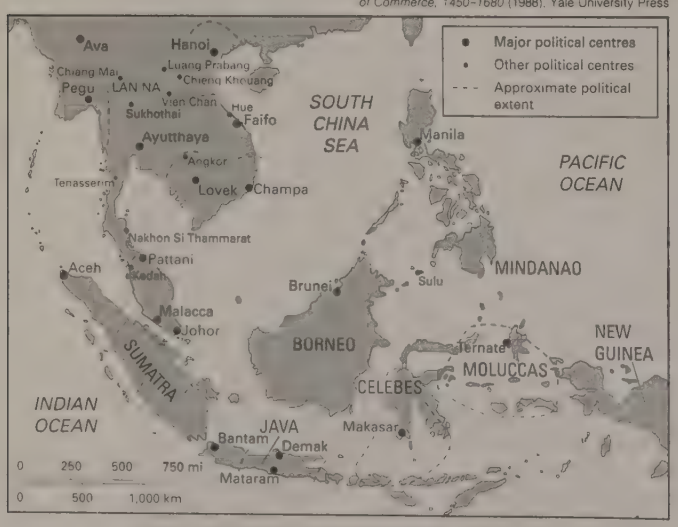
State and society. There were five major powers in Southeast Asia between the 14th and 18th centuries: Myanmar under the rulers of Ava (1364–1752), especially the Toungoo dynasty during most of that period; an independent Vietnam under the Later Le dynasty (1428–1788); the Tai state of Ayutthaya, or Ayudhia (1351–1767); Majapahit, centred on Java (1292–*c.* 1527); and Malacca (Melaka) centred on the Malay Peninsula (*c.* 1400–1511). Particularly with the waning of Indian influence (the last known Sanskrit inscription dates from the late 13th century), each power had developed in distinctive ways: more than ever, what constituted being "Javanese" or "Burman," for example, was taking focus, and the Vietnamese, too, sought to clarify what was their own as opposed to what was Chinese. Remarkably enough, the process by which this was accomplished was characterized not by elimination or purification but by absorption. The syncretic powers developed in earlier periods had by no means weakened. The Tai, comparative newcomers, absorbed much of Khmer civilization during this period and, beginning with their written language, shaped it to their requirements. The Burmans absorbed Mon civilization in a similar fashion, and the Javanese of Majapahit could not help but make adjustments with the Malay and other cultures of the archipelago that they came to dominate. Even the Vietnamese, who had decided after several generations of struggle to adopt the outlines of a Confucian state that they had inherited from China, in the late 14th and early 15th centuries not only modified that model but also absorbed important influences from the culture of the Cham, an Indianized people whose kingdom, Champa, they had decisively (though not finally) defeated in 1471. This integrative approach may not have represented a conclusive departure from the behaviour of the ancient *mandala* states, but it does seem to have sustained larger and more far-reaching states, as well as richer and more complex elite cultures.

At the same time, however, a galaxy of smaller states appeared, some of them very powerful for their size and all of them ambitious. These states were especially numerous in insular Southeast Asia, where Aceh, Bantam (Banten), Makasar (Macassar), and Ternate were only the most prominent of many such Islāmic sultanates; on the mainland, Chiang Mai (Chiengmai), Luang Prabang, and Pegu at various times during the period were powerful enough to be taken seriously. They both imitated and contributed to the court cultures of their larger neighbours and made alliances, war, and peace with many powers.

Rise of trade

Above all, these states participated in a dynamic and prosperous trade, not merely in exotics or high-value goods (such as gems and metal items) but in such relatively mundane goods as salted dried fish, ceramics, and rice.

Adapted from A.J.S. Reid *Southeast Asia in the Age of Commerce, 1450–1680* (1988), Yale University Press



Major political centres of Southeast Asia, *c.* 1600.

While institutions of servitude were structured somewhat differently from those of the West, there was no mistaking that a lively trade in human beings prized for their labour or craftsmanship took place. The proliferation of states and the rapid growth of an accompanying intricate web of local cultural and commodity exchange laid the foundation for both greater local autonomy and increased regional interdependency.

The dynamics of regional trade brought change to most Southeast Asian societies during this period. These changes were by no means uniform; the effect on hill tribes subject to periodic raiding, for example, was understandably different from that on coastal communities suddenly wealthy from trade. In some instances the alterations must have been dramatic: the native sago diet of many inhabitants of the Moluccas (Maluku) region, for example, was displaced by one based on rice brought from Java, more than 1,500 miles to the west. Yet it does seem that some changes were felt widely, especially in the larger states. Perhaps the most important was that, while old ideas of kingship and sovereignty were cultivated, in reality much power—and in some places critical power—had fallen into the hands of a merchant class. The royal courts themselves often dabbled in trade to an unprecedented degree. It perhaps is not accurate to say that kingship as an institution was weakening, but the courts, particularly in insular Southeast Asia, became more complicated centres of elite power.

Urbanization was another development of importance. Although some societies, notably that of the Javanese, seem not to have been affected, the growth of large and densely populated centres was a widespread phenomenon. By the 16th century some of these rivaled all but the very largest European cities. Malacca, for example, may have had a population of 100,000 (including traders) in the early 16th century; in Europe only Naples, Paris, and perhaps London were larger at that time. Finally, Southeast Asians during the 16th and 17th centuries appear to have enjoyed good health, a varied diet, and a comparatively high standard of living, especially when compared with most of the population of Europe of the same period.

Religion and culture. New religions appeared in Southeast Asia, accompanying the currents of trade and often entwined with social changes already underway. Gradually, in most areas, these religions filled the gaps left by weakening local Hindu-Buddhist establishments and beliefs, and by the mid-18th century the region had assumed something much like its modern religious configuration. On the mainland, Theravāda Buddhism, which had been making inroads in Cambodia since the 11th century, underwent revitalization, the result especially of royal patronage and direct contact with Theravāda monasteries in Sri Lanka. Both the general idiom and many precepts of Theravāda already were familiar in Indianized societies, making this a gentle, nearly silent revolution that despite its subtlety was no less important. In Ayutthaya and the other Tai kingdoms and in the Mon-Burman states, Theravāda Buddhism buoyed the kingship and introduced a vigorous intellectual leadership; it also spread broadly among the populace and thus played an important role as a cohesive social and cultural force from which the people of modern Thailand and Myanmar later were to draw much of their sense of identity.

Christianity

Christianity made its appearance in the early 16th century, brought by the Portuguese, Spanish, and, somewhat later, the French. It spread easily in the northern Philippines, where Spanish missionaries did not have to compete with an organized religious tradition and could count on the interested support of a government bent on colonization. Unlike the religions with which Southeast Asia had been familiar, Christianity showed no interest in syncretic accommodation of local animist or other beliefs. The Spanish friars rooted out whatever they could find in the way of indigenous tradition, destroying much of cultural value, including, it appears, a native writing system. By the 18th century, most of the Philippines, except the Muslim south, was Roman Catholic, and a society that was both Filipino and Christian had begun to evolve. Elsewhere in Southeast Asia, however—with the exception of Vietnam and parts of the Moluccas island

group of eastern Indonesia—Christianity attracted little interest. It did not go unopposed and was resisted, for example, by Buddhist monks in Thailand and Cambodia in the 16th century, but Christian doctrines do not appear to have attracted the general populace. There were few conversions, and rulers were not unduly disturbed by the presence of missionaries, except on occasions when they were accompanied by political and economic adventurers; these people were crushed.

Islām, however, captured the imagination of Southeast Asians in the archipelago. It was proselytized primarily by Malacca and Aceh after 1400 and by the late 17th century was the dominant faith from the western tip of Sumatra to the Philippine island of Mindanao. The conversion process was gradual, for Muslim traders from the Middle East and India long had traveled the sea route to China; it seems likely that they traded and settled in the port cities of Sumatra and Java as early as the 9th or 10th century. Perhaps as a result of weakening of the Hindu-Buddhist courts and the rise of smaller, independently minded trading states and social classes, Islām made important inroads among both ruling elites and others.

Conversion was comparatively easy and promised certain practical advantages, especially in trade, to members of the Islāmic community (the *ummah*). In addition, Islām was itself diverse, offering a spectrum of approaches from mystical to fundamentalist, and in practice Muslim proselytizers often were tolerant of syncretic behaviour. In addition, Islāmic culture, especially poetry and philosophy, was particularly attractive to courts anxious to enhance their status as cultural hubs. While the spread of Islām throughout the archipelago was not entirely peaceful, for the most part it proceeded in evolutionary fashion and without remarkable disturbance. Javanese Muslims, perhaps even members of the court, lived peacefully in the capital of Hindu-Buddhist Majapahit, for example, and Muslims and non-Muslims everywhere continued to trade, enter into alliances, and inhabit the same general cultural world. What change there was tended to occur slowly in the face of robust and deeply rooted tradition. In some societies the cultural response was original and lively. Along the northern coast of Java, for example, architecture, batik cloth-dyeing motifs, and the literature and performance of the wayang (shadow-puppet theatre) were deeply affected by Islāmic ideas and produced vital new forms to accompany the old.

Chinese and Western incursions. Southeast Asia, unlike many other parts of the world on the eve of European expansion, long had been a cosmopolitan region acquainted with a diversity of peoples, customs, and trade goods. The arrival of Europeans in force in the early 16th century (others had made visits earlier, beginning with Marco Polo in 1292) caused neither wonderment nor fear. Long-distance travel by then was no novelty, and already there was impressive precedence for the arrival of foreign delegations rather than of individual trading vessels. A century before the Portuguese first arrived at Malacca in 1509, that port and a number of others in Southeast Asia had been visited by a succession of Chinese fleets. Between 1403 and 1433 Ming-dynasty China had sent several enormous flotillas of as many as 63 large vessels and up to 30,000 people on expeditions that carried them as far as Africa. The purpose of these journeys, led by the Muslim court eunuch Cheng Ho, was to secure diplomatic and trade advantages for the Chinese and to extend the sovereign lustre of the ambitious Yung-lo Emperor. Yet, except for efforts to regain Dai Viet (Vietnam) as a province, these expeditions had no permanent military or colonial ambitions and did not much disturb the Southeast Asian region. Perhaps in part because of the sound defeat the Vietnamese handed a Ming occupying army in 1427, China lost interest in its new and far-flung initiatives, and the voyages came to an abrupt end.

Voyages of Cheng Ho

Europeans presented a rather different prospect for Southeast Asia, however, above all because they sought riches and absolute control over the sources of this wealth. The Europeans were few in number, often poorly equipped, and generally could not claim great technological superiority over Southeast Asians, but they were also determined, often well-organized and highly disciplined fighters, and utterly ruthless and unprincipled. Except for the Spanish in the Philippines, they were not interested in colonization but rather in the control of trade at the lowest financial cost. These characteristics made Europeans a formidable—though by no means dominant—new force in Southeast Asia. Except in a few locales and special circumstances, for the better part of 250 years Europeans could accomplish little politically or militarily without strong Southeast Asian allies. Individual adventurers often were useful to a particular Southeast Asian ruler or aspirant to the throne, but they were carefully watched and, when necessary, dispatched. Constantine Phaulkon, the Greek advisor to the Siamese court who was executed in 1688 on charges of treason, was only the most dramatic example. In economic affairs, Europeans soon discovered that they were quite unable, even by the most drastic means, to monopolize the spice trade for which they had come. They generally were forced to engage in commerce by Southeast Asian rules and soon found themselves dependent on the local carrying trade for survival. For these reasons, the celebrated Portuguese conquest of Malacca in 1511 did not signal the dawn of an age of Western dominance in Southeast Asia. The majority of the population and much of the trading activity deserted the port, the sultan moved his court elsewhere, and by the end of the 16th century Malacca was a backwater; the Malay trade flourished elsewhere into the 18th century.

Yet it would be a mistake to conclude that the Western presence represented nothing more than a minor irritant. European commercial tools, especially the ability to amass large amounts of investment capital, were different and, from a capitalistic point of view, more sophisticated and dynamic than those of the Southeast Asians. The Dutch and British East India companies often were able to make inroads on certain markets simply by having a large amount of money available, and it was possible for them to adopt long-term strategies by carrying large deficits and debts. Although company directors in Europe warned against the dangers—and costs—of involvement in local affairs, the representatives on the spot often could see no other course. Thus, soon after permanently establishing themselves on Java in 1618, the Dutch found themselves embroiled in the succession disputes of the court of Mataram and, by the late 1740s, virtual kingmakers and shareholders in the realm. Finally, Europeans did bring with them much that was new. Some items shaped Southeast Asian life in unexpected ways: the chili pepper, which the Spanish introduced from the New World, came to hold such an important place in the region's diet that today Southeast Asian cuisine can hardly be imagined without it. Another import, however, was coffee, with a more ominous effect. Smuggled into Java in 1695 against Dutch East India Company rules, coffee by the early 18th century had become a company monopoly produced through a unique relationship between the Dutch and the local Javanese elite in a system that prefigured the one adopted by the 19th-century colonial state.

Introduction of coffee

PATTERNS OF A COLONIAL AGE

Crisis and response. In the last half of the 18th century, all the major states of Southeast Asia were faced with crisis. The great political and social structures of the classical states had begun to decay, and, although the reasons for this disintegration are not altogether clear, the expanded size of the states, the greater complexity of their societies, and the failure of older institutions to cope with change all must have played a part. It is also likely that European efforts to choke and redirect the region's trade had already done much to destroy the general prosperity that trade previously had provided, though Europeans were neither ubiquitous nor in a position to rule, even in Java. The most serious circumstances were undoubtedly those of Vietnam, where from 1771 to 1802 there raged a struggle—the Tay Son rebellion—over the very nature of the state. This rebellion threatened to sweep away the entire Confucian establishment of Vietnam, and perhaps would have done so if its leader had not attempted to accomplish too much too quickly. Elsewhere, war and

confusion held societies in their grip for much shorter periods, but everywhere rulers were compelled to think of changed circumstances around them and what they meant for the future.

In the mainland states three great rulers of three new dynasties came to the fore: Bodawpaya (ruled 1782–1819) in Myanmar, Rama I (1782–1809) in Siam (Thailand), and Gia Long (1802–20) in Vietnam. All three were fully aware of the dangers, internal as well as external, that faced them and their people, and their efforts were directed at meeting these challenges. As their armies extended their reach beyond earlier limits, these rulers vigorously pursued a combination of traditional and new policies designed to strengthen their realms. Of particular importance were efforts to bring villages under closer state control, curb shifting patron-client relationships, and centralize and tighten the state administrative apparatus. The institution of kingship itself seemed to become more dynamic and intimately involved in the direction of the state. In retrospect, some of these policies had a recognizably modern ring to them, and taken together they represented, if not a revolution, at least a concerted effort at change. Even Gia Long, whose conscience and circumstance both demanded that he give special attention to reviving the classical Confucian past, quietly incorporated selected Western and Tay Son ideas in his government. Nor were the changes ineffectual, for by 1820 the large mainland states stood at the height of their powers. Nevertheless, it was uncertain whether these efforts would be sufficient to withstand the pressures of the immediate future.

In insular Southeast Asia the Javanese state confronted a similar crisis, but it had far less freedom with which to respond. The Gianti Agreement (1755) had divided the realm and given the Dutch decisive political and economic powers. Though resistance was not impossible, it was difficult, especially since the rulers and their courts were now largely beholden to the Dutch for their positions. The elite's response to these circumstances generally has been interpreted as a kind of cultural introversion and avoidance of reality, a judgment that probably is too harsh. The Javanese culture and society of earlier days was no longer serviceable, and court intellectuals sought to find a solution in both a revitalization of the past and a clear-eyed examination of the present. Neither effort was successful, though not for want of trying. The idea of opposing Dutch rule, furthermore, was not abandoned entirely, and it was only the devastating Java War (1825–30) that finally tamed the Javanese elite and, oddly enough, left the Dutch to determine the final shape of Javanese culture until the mid-20th century.

The Java War

Western dominance. Except in Java and much of the Philippines, the expansion of Western colonial rule in most of Southeast Asia was a phenomenon only of the 19th and the beginning of the 20th centuries. In the earlier period Europeans tended to acquire territory as a result of complicated and not always desired entanglements with Southeast Asian powers, either in disputes or as a result of alliances. After about 1850, Western forces generally were more invasive, requiring only feeble justification for going on the attack. The most important reasons for the change were a growing Western technological superiority, an increasingly powerful European mercantile community in Southeast Asia, and a competitive scramble for strategic territory. Only Siam remained largely intact and independent. By 1886 the rest of the region had been divided among the British, French, Dutch, and Spanish (who soon were replaced by the Americans), with the Portuguese still clinging to the island of Timor. What were often called "pacification campaigns" were actually colonial wars—notably in Burma (Myanmar), Vietnam, the Philippines, and Indonesia—and continued well into the 20th century. More peaceful Western encroachments on local sovereignty also occurred until the 1920s. Full-blown, modern colonial states existed for only a short period, in many cases for not much more than a generation.

These colonial regimes, however, were not insubstantial, as they put down strong bureaucratic roots and—though often co-opting existing administrative apparatuses—formed centralized, disciplined structures of great power. They were backed by the enormous economic resources of the industrialized Western nations; and by the early 20th century, having effectively disarmed the indigenous societies, they possessed a monopoly on the means of violence. There is no mistaking the impact of Western colonial governments on their surroundings, and nowhere is this more evident than in the economic sphere. Production of tin, oil, rubber, sugar, rice, tobacco, coffee, tea, and other commodities burgeoned, driven by both government and private activity; this brought rapid changes to the physical and human landscape and coupled Southeast Asia to a new worldwide capitalist system.

Indeed, colonial domination was only a variant condition in a rapidly changing world. Siam, which through a combination of circumstance and the wise leadership of Mongkut (ruled 1851–68) and Chulalongkorn (1868–1910) avoided Western rule, nevertheless was compelled to adopt policies similar to, and often even modeled on, those of the colonial powers in order to survive. Modernization appeared to require such an approach, and the Thai did not hesitate to embrace it with enthusiasm. Bangkok in the late 1920s surpassed even British Singapore as a centre of such modern amenities as electric lighting and medical facilities, and the state itself had achieved an enviable degree of political and economic viability among its colonial neighbours. The Thai may have "colonized themselves," as some critics have noted, but in so doing they also escaped or diluted some of the more corrosive characteristics of Western rule, among them racism and cultural destruction. They also do not appear to have experienced the same degree of rural unrest that troubled their colonial neighbours in the 1920s and '30s. They were unable, however, to avoid other concomitants of state expansion and modernization.

Siamese response

Transformation of state and society. It was not the purpose of the new states to effect rapid or broad social change. Their primary concerns were extending bureaucratic control and creating the conditions for success in a capitalist world economy; the chief necessity was stability or, as the Dutch called it, *rust en orde* ("tranquillity and order"). Boundaries were drawn, villages defined, laws rewritten—all along Western lines of understanding, often completely disregarding indigenous views and practices—and the new structure swiftly replaced the old. Social change was desired only insofar as it might strengthen these activities. Thus, the Thai began early on to send princes to Europe for their education, employing them throughout the government on their return. The Dutch created exclusive schools for the indigenous administrative elite—a kind of petty royalty—and invented ways of reducing social mobility in this group, as, for example, by making important positions hereditary. But the new governments did not provide Western-style learning to most Southeast Asians, primarily because it was an enormous, difficult, and expensive task and also because policymakers worried about the social and political consequences of creating an educated class. Except in the Philippines, by the mid-1930s only a small percentage of indigenous children attended government-run schools, and only a fraction of those studied above the primary-school level. Some Southeast Asian intellectuals soon drew the conclusion that they had better educate themselves, and they began establishing their own schools with modern, secular courses of study. Some, like the Tonkin Free School in Vietnam (1907), were closed by the colonial regimes, their staffs and pupils hounded by police; others, like the many so-called "wild schools" in Indonesia in the 1930s, were much too numerous to do away with altogether, but they were controlled as carefully as possible.

Nevertheless, during the 1920s and '30s a tiny but thoughtful and active class of Westernized Southeast Asian intellectuals appeared. They were not the first to literally and figuratively speak the language of the colonial rulers and criticize them, for by the turn of the 20th century Java and Luzon, with the longest experience under Western rule, had already produced individuals like the Javanese noblewoman Raden Adjeng Kartini and the Filipino patriot José Rizal. The newer generation, however, was more certain in its opposition to colonial rule (or, in Siam,

Rise of nationalist leaders

rule by the monarchy), clearer and far more political in its conception of a nation, and unabashedly determined to seize leadership and initiative in their own societies. In Burma this group called themselves *thakin* (Burmese: "master"), making both sarcastic and proud use of an indigenous word that had been reserved for Burmese to employ when addressing or describing Europeans. These new intellectuals were not so much anti-Western as they were anticolonial. They accepted the existing state as the foundation of a modern nation, which they, rather than colonial officials, would control. This was the generation that captained the struggles for independence (in Siam, independence from the monarchy) and emerged in the post-World War II era as national leaders. The best-known figures are Sukarno of Indonesia, Ho Chi Minh of Vietnam, and U Nu of Burma (subsequently Myanmar).

The chief problem facing the new intellectuals lay in reaching and influencing the wider population. Colonial governments feared this eventuality and worked to prevent it. Another obstacle was that the ordinary people, especially outside cities and towns, inhabited a different social and cultural world from that of the emerging leaders. Communication was difficult, particularly when it came to explaining such concepts as nationalism and modernization. Still, despite Western disbelief, there was considerable resentment of colonial rule at the lower levels of society. This was based largely on perceptions that taxes were too numerous and too high, bureaucratic control too tight and too prone to corruption, and labour too coercively extracted. In many areas there also was a deep-seated hatred of control by foreigners, whether they be the Europeans themselves or the Chinese, Indians, or others who were perceived as creatures of their rule. Most of the new intellectual elite were only vaguely aware of these sentiments, which in any case frequently made them uneasy; in a sense they, too, were foreigners. In the 1930s, however, a series of anticolonial revolts took place in Burma, Vietnam, and the Philippines; though they failed in their objectives, these revolts made it clear that among the masses lay considerable dissatisfaction and, therefore, radical potential. The revolts, and the economic disarray of the Great Depression, also suggested that European rule was neither invulnerable nor without flaws. When the outbreak of war in Europe and the Pacific showed that the colonial powers were much weaker militarily than had been imagined, destroying colonial rule and harnessing the power of the masses seemed for the first time to be real possibilities.

Japanese occupation. The arrival of the Japanese armed forces in Southeast Asia in 1941–42 did not, however, occasion independence. A few leaders perhaps had been naive enough to think that it might—and some others clearly admired the Japanese and found it acceptable to work with them—but on the whole the attitude of intellectuals was one of caution and, very quickly, realization that they were now confronted with another, perhaps more formidable and ferocious, version of colonial rule. The Japanese had no plans to radicalize or in any way destabilize Southeast Asia—which, after all, was slated to become part of a Tokyo-centred Greater East Asia Co-prosperity Sphere; in the short term they sought to win the war, and in the long run they hoped to modernize the region on a Japanese model. Continuity served these purposes best, and in Indochina the Japanese even allowed the French to continue to rule in return for their cooperation. Little wonder that before long Southeast Asians began to observe that, despite "Asia for the Asians" propaganda, the new and old colonial rulers had more in common with each other than either had with the indigenous peoples.

Difference from Western rule

Still, for two distinct reasons the period does represent a break from the past. First, the Japanese attempted to mobilize indigenous populations to support the war effort and to encourage modern, cooperative behaviour on a mass scale; such a thing had never been attempted by Western colonial governments. Virtually all of the mobilization efforts, however, were based on Japanese models, and the new rulers were frustrated to discover that Southeast Asians did not behave in the same fashion as Japanese. Frequently the result was disorder, corruption, and, by the end of the war, a seething hatred of the Japanese. It was also the case that, both because the war was going against them and because the response to other approaches was unenthusiastic, the Japanese were compelled before long to utilize local nationalism in their mobilization campaigns, again something quite impossible under European rule. The consequences were to benefit local rather then Japanese causes and, ironically, to contribute handsomely to the building of anti-Japanese sentiments.

A second difference between Western and Japanese colonialism was in the opportunities the occupation provided the new educated elite. The Japanese were wary of these people because of their Western orientation but also favoured them because they represented the most modern element in indigenous society, the best partner for the present, and the best hope for the future. Often dismissed as "pseudo-intellectuals" by the Western colonial governments and prevented from obtaining any real stake in the state, the new intellectuals under the Japanese were accorded positions of real (though not unlimited or unsupervised) authority. Nor could Southeast Asians who found themselves in these positions easily fault the policies they now accepted responsibility for carrying out or at least supporting, since many of these policies were in fact—if not always in spirit—similar to ones they had endorsed in earlier decades. In short, the Western-educated elite emerged from the Japanese occupation stronger in various ways than they had ever been. By August 1945 they stood poised to inherit (or, given the variety of political conditions at the end of the war, to struggle among themselves over inheriting) the mantle of leadership over their own countries.

Southeast Asia was changed in an evolutionary, rather than revolutionary, way by the Japanese occupation. Although returning Europeans and even some Southeast Asians themselves complained that Japanese fascism had deeply influenced the region's societies, there is not much evidence that this was the case. Japanese rule, indeed, had destroyed whatever remained of the mystique of Western supremacy, but the war also had ruined any chances that it might be replaced with a Japanese mystique. There was clearly little clinging to Japanese concepts except where they could be thoroughly indigenized; even the collaboration issue, so important to Europeans and their thinking about the immediate postwar era, failed to move Southeast Asians for long. And, if the general population appeared less docile in 1945 than four years earlier, the reason lay more in the temporary removal of authority at the war's end than in the tutelage of the Japanese.

CONTEMPORARY SOUTHEAST ASIA

Postwar events

Struggle for independence. The swift conclusion of the war in the Pacific made it impossible for the former colonial masters to return to Southeast Asia for several weeks, in some areas for months. During the interim, the Japanese were obliged by the Allies to keep the peace, but real power passed into the hands of Southeast Asian leaders, some of whom declared independence and attempted with varying degrees of success to establish government structures. For the first time since the establishment of colonial rule, firearms in large numbers were controlled by Southeast Asians. Such was the groundwork for the establishment of new, independent states.

Prewar nationalism had been most highly developed in Vietnam and Indonesia, and the colonial powers there were least inclined to see the new realities created by the war, perhaps because of the large numbers of resident French and Dutch and because of extensive investments. The result in both countries was an armed struggle in which the Western power was eventually defeated and independence secured. The Indonesian revolution, for all its internal complexities, was won in little more than four years with a combination of military struggle and civilian diplomacy. The revolution of the Vietnamese, who had defeated the French by 1954, continued much longer because of an internal political struggle and because of the role Vietnam came to play in global geopolitics, which ultimately led to the involvement of other external powers, among them the United States. In both cases, however, independence

was sealed in blood, and a mythologized revolution came to serve as a powerful, unifying nationalist symbol. In the rest of Southeast Asia, the achievement of independence was, if not entirely peaceful, at least less violent. Malaysia and the Philippines suffered "emergencies" (as armed insurgencies were euphemistically called), and Burma, too, endured sporadic internal military conflict. For better or worse, these conflicts were no substitutes for a genuine revolutionary experience.

Whether by revolution or otherwise, decolonization proceeded rapidly in Southeast Asia. The newly independent states all aspired toward democratic systems more or less on the Western model, despite the lack of democratic preparation and the impress of nationalist sentiment. None expressed a desire to return to precolonial forms of government, and, although some Western observers professed to see in such leaders as Indonesia's Sukarno Southeast Asian societies returning to traditional behaviour, their judgment was based more on ephemeral signs than on real evidence. For one thing, societies as a whole had been too much altered in the late 19th and early 20th centuries to make it clear what "tradition" really was. For another, the new leadership retained the commitment to modernization that it had developed earlier. They looked forward to a new world, not an old one. The difficulty, however, was that there was as yet little consensus on the precise shape this new world should take, and colonial rule had left indigenous societies with virtually no experience in debating and reaching firm decisions on such important matters. It is hardly surprising that one result of this lack of experience was a great deal of political and intellectual conflict. Often forgotten, however, is another result: an outpouring of new ideas and creativity, particularly in literature. This signaled the beginning of a kind of cultural renaissance, the dimensions and significance of which are still insufficiently understood.

Early independence period

Defining new states and societies. The first two decades of independence constituted a period of trial and error for states and societies attempting to redefine themselves in contemporary form. During this time, religious and ethnic challenges to the states essentially failed to split them, and (except in the states of former Indochina) both communism and Western parliamentary democracy were rejected. Indonesia, the largest and potentially most powerful nation in the region, provided the most spectacular examples of such developments, ending in the tragic events of 1965–66, when between 500,000 and 1,000,000 lives may have been lost in a conflict between the Indonesian Communist Party and its opponents. Even Malaysia, long the darling of Western observers for its apparent success as a showcase of democracy and capitalist growth, was badly shaken by violence between Malays and Chinese in 1969. The turmoil often led Southeast Asia to be viewed as inherently unstable politically, but from a longer perspective—and taking into account both the region's great diversity and the arbitrary fashion in which boundaries had been set by colonial powers—this perhaps has been a shortsighted conclusion.

The new era that began in the mid-1960s had three main characteristics. First, the military rose as a force in government, not only in Vietnam, Burma, and Indonesia but also in the Philippines and—quietly—in Malaysia. The military establishments viewed themselves as actual or potential saviours of national unity and also as disciplined, effective champions of modernization; at least initially, they frequently had considerable support from the populace. Second, during this period renewed attention was given by all Southeast Asian nations to the question of unifying (secular and national) values and ideology. Thailand, Indonesia, and Vietnam had been first in this area in the 1940s and '50s, but the others followed. Even Singapore and Brunei developed ideologies, with the express purpose of defining a national character for their people. Finally, virtually all Southeast Asian states abandoned the effort of utilizing foreign models of government and society—capitalist or communist—and turned to the task of working out a synthesis better suited to their needs and values. Each country arrived at its own solution, with varying degrees of success. By the 1980s what generally had emerged were quasi-military bourgeois regimes willing to live along modified democratic lines—*i.e.*, with what in Western eyes appeared to be comparatively high levels of restriction of personal, political, and intellectual freedom. Whatever their precise political character, these were conservative governments. Even Vietnam, the most revolutionary-minded among them, could not stomach the far-reaching and murderous revolution of the Khmer Rouge in Cambodia in the mid-1970s and by the end of the decade had moved to crush it.

Tempting as it may be to conclude that greater doses of authoritarian rule (some of it seemingly harking back directly to colonial times) merely stabilized Southeast Asia and permitted the region to get on with the business of economic development, this approach was not successful everywhere. In Burma (called Myanmar since 1989) the military's semi-isolationist, crypto-socialist development schemes came to disaster in the 1980s, revealing the repressive nature of the regime and bringing the country to the brink of civil war by the end of the decade. In the Philippines the assault by President Ferdinand Marcos and his associates on the old ruling elite class brought a similar result, in addition to a spectacular level of corruption and the looting of the national treasury. In Vietnam, where the final achievement of independence in 1975 brought bitter disappointment to many and left the country decades behind the rest of the region in economic development, public and internal Communist Party unrest forced an aging generation of leaders to resign and left the course for the future in doubt as never before.

Failures in authoritarianism

The states generally thought to be most successful to date—Thailand, Indonesia, Malaysia, and especially Singapore—have followed policies generally regarded as moderate and pragmatic. All are regarded as fundamentally stable and for that reason have attracted foreign aid and investment; all have achieved high rates of growth since the mid-1970s and enjoy the highest standards of living in the region. Their very success, however, has created unexpected social and cultural changes. Prosperity, education, and increasing access to world media and popular culture have all given rise, for example, to various degrees of dissatisfaction with government-imposed limitations on freedom and to social and environmental criticism. Particularly in Indonesia and Malaysia, there has been a noticeable trend toward introspection and discussion of national character, as well as a religious revival in the form of renewed interest in Islām. It appears that the comparatively small and unified middle class, including a generally bureaucratized military, is becoming larger, more complex, and less easily satisfied. That was undoubtedly not the intent of those who framed governmental policy, but it is a reality with which they must deal.

Reappearance of regional interests. After the end of the 17th century, the long-developed polities of Southeast Asia were pulled into a Western-dominated world economy, weakening regional trade networks and strengthening ties with distant colonial powers. In the early years of independence these ties often remained strong enough to be called neocolonial by critics, but after the mid-1960s these partnerships could no longer be controlled by former colonial masters, and the new Southeast Asian states sought to industrialize and diversify their markets. On the one hand, this meant a far greater role for Japan in Southeast Asia; that country is by far the most important trading partner of most Southeast Asian nations. On the other, it meant that many countries began to rediscover commonalities and to examine the possibilities within the region for support and markets.

Formation of ASEAN

In 1967 the Association for Southeast Asian Nations (ASEAN) was formed by Malaysia, Indonesia, the Philippines, Thailand, and Singapore (Brunei joined in 1985). This group's initial interest was in security, but it has moved cautiously into other fields. It played an important role, for example, in seeking an end to the Vietnam-Cambodia conflict and has sought a solution to the civil strife in Cambodia. In economic affairs it has worked quietly to discuss such matters as duplication of large industrial projects, but, perhaps because the economies of most of its members are quite similar and as yet only par-

tially industrialized, ASEAN has not attempted to build a true economic community. Only since the mid-1980s has ASEAN been taken seriously by major powers, or even sometimes by Southeast Asians themselves. It seems likely, however, that the formerly Soviet-dominated states of Vietnam, Laos, and Cambodia will become part of ASEAN before the end of the 1990s, and Myanmar may be compelled to follow. Such circumstances will undoubtedly open up greater regional markets and give the region as a whole a more imposing world profile. Moreover, modern communications, which have already begun to inform ASEAN populations more closely about each other, cannot help but further this process and draw attention to common strands in an emerging modern culture that is shared, at least to some degree, by all the nations of the region. (W.H.F.)

THE COUNTRIES OF SOUTHEAST ASIA

Brunei

Brunei—officially State of Brunei, Abode of Peace (Malay: Negara Brunei Darussalam)—is a small, independent Islāmic sultanate on the northern coast of the island of Borneo. It has an area of 2,226 square miles (5,765 square kilometres) and is bounded to the north by the South China Sea and on all other sides by the East Malaysian state of Sarawak, which also divides the state into two enclaves of unequal size. The western enclave is the larger of the two and contains the capital city of Bandar Seri Begawan. Brunei achieved independence in 1984, after having been a British protectorate since 1888. It is a member of the Commonwealth and the Association of Southeast Asian Nations (ASEAN).

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief, drainage, and soils.* The narrow coastal plain in the north gives way to rugged hills farther south. Brunei's highest point is Pagon Peak (6,070 feet [1,850 metres]), in the southeast. The country is drained by the Belait, Tutong, and Brunei rivers in the western enclave and by the Pandaruan and Temburong rivers in the east; all flow generally northward to the South China Sea. The Belait is the largest river in the country. The soils of Brunei are deeply weathered and highly leached and generally are infertile. Richer alluvial soils are found along the rivers and in some parts of the coastal floodplain, and these offer the best agricultural potential.

Climate. The climate of Brunei is governed by the equatorial monsoon winds. Daily temperatures average between 76° and 86° F (24° and 30° C). Precipitation averages 110 inches (2,800 millimetres) annually in the coastal areas but can exceed 150 inches farther inland. Rainfall is heaviest during the northeast monsoon (December to March), with lesser amounts falling during the southwest monsoon (May to October).

The monsoons

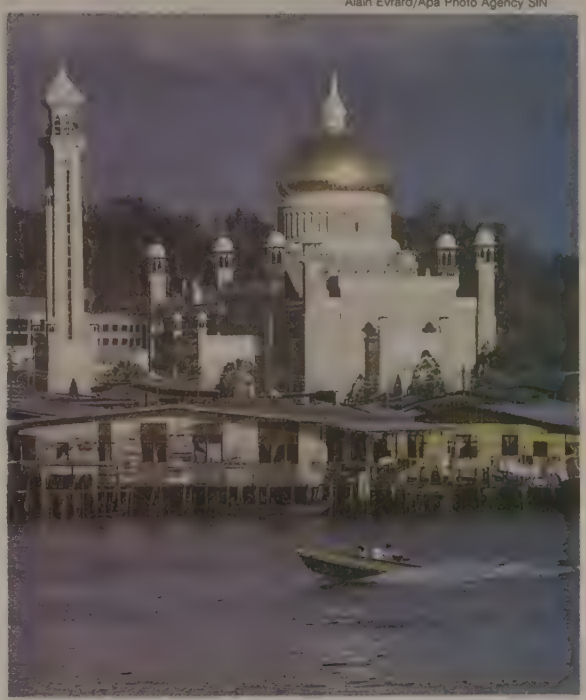
Plant and animal life. About three-fifths of the country is covered with virgin tropical rain forest, with another one-fifth under second-growth forest. The undisturbed rain forest consists mainly of hardwoods of the Dipterocarpaceae family, most of which are of commercial importance. Large expanses of peat and freshwater swamps are found in the poorly drained lowlands of the Belait and Tutong rivers, while mangrove swamps are common along the lower riverine reaches and sheltered coastal areas. The complex vegetation of the rain forest provides niches for a rich variety of animals, including monkeys, birds, reptiles, and insects.

The people. About two-thirds of the population of Brunei is Malay, and nearly one-fifth is Chinese. The remainder includes such indigenous peoples as the Kadazan (Kedazan), Murut, Bisaya (Bisayah), and Iban (or Sea Dayak), as well as small numbers of Indians and foreigners. The official language is Malay, with English as a major second language. Brunei's population is predominantly Sunnite Muslim, although the Chinese usually are Buddhist, Taoist, or Confucian. Some of the indigenous peoples are Christian, while others in the remote interior are spirit worshipers.

The interior upland areas are sparsely populated by indigenous peoples, who clear areas of the forest for shifting cultivation. In the western enclave, the Iban inhabit the westernmost Belait district, and the Kadazan live in the rural areas of the central Tutong district; the Murut and Bisaya have settled mainly in the eastern enclave. The Malays are distributed in the riverine and coastal villages and towns, and the Chinese are concentrated in the urban areas.

About two-thirds of Brunei's population is found in and around Bandar Seri Begawan, while one-fifth lives in the oil-rich coastal areas of the Belait district. The capital, located on the Brunei River about nine miles from its mouth on Brunei Bay, is the largest urban centre. Adjacent to the modern section of the city is an older part called Kampong Ayer, where Brunei Malays live in houses built on stilts along inlets of the river.

Alain Evrard/Apa Photo Agency SIN



The Sultan Omar Ali Saifuddin Mosque rising above Kampong Ayer (foreground), Bandar Seri Begawan, Brunei.

The economy. Brunei's economy is almost totally dependent on the exploitation of its vast reserves of petroleum and natural gas. Oil and gas revenues have allowed the state to give its citizens one of the highest per capita incomes in Asia, but it also has made the country dependent on a single commodity that is subject to market fluctuations. In addition, Brunei must rely exclusively on imports for nearly all of its manufactured goods and most of its food.

Resources. In addition to its reserves of petroleum and natural gas, Brunei has rich—though undeveloped—deposits of white-quartz sand. Oil was first produced in 1929, while the natural gas industry was developed after the discovery in the 1960s of large deposits. Nearly all the oil and natural gas produced in Brunei comes from offshore fields located off the western enclave. All but a small percentage of the production is exported, a small refinery supplying local needs. Oil output peaked in the late 1970s and subsequently was reduced in order to conserve reserves. Brunei's huge deposits of natural gas were intensively developed in the 1970s, including the construction

Hydrocarbon production

of a gas liquefaction plant. As a result, export earnings from liquid natural gas have come to roughly equal those from petroleum.

Agriculture, fishing, and forestry. These three activities—once the mainstays of Brunei's economy—now constitute only a tiny and dwindling fraction of the gross domestic product. The proportion of the workforce practicing agriculture and fishing has fallen, as people have abandoned these occupations in search of better-paying jobs; production in both areas fails to meet local demand. Rubber, once a major crop, is no longer cultivated. The area planted in rice has declined, although the government has tried to reverse that trend by encouraging production. Other crops—mainly sago, coconuts, fruits, and vegetables—are grown on a small scale. The timber industry has been developed to supply the local market exclusively, while the export of timber has been prohibited.

Finance and trade. Brunei's revenues from petroleum and natural gas, which constitute nearly all of its export earnings, have resulted in perennial trade surpluses. Much of the state's financial activity is concerned with managing its substantial foreign investments, and the return on these has become an important source of income. Brunei's principal trading partners are Japan, the members of the European Community, Singapore, and South Korea.

Transportation. Brunei traditionally depended on its rivers and the sea for transportation. Rivers have remained the main means of transport into the interior, but a good and rapidly expanding network of roads has been built in the coastal areas; some roads have been extended into the interior. Per capita car ownership in Brunei is one of the highest in the world. Brunei has two major ports: a large, deepwater harbour at Muara, on Brunei Bay, and a smaller port at Kuala Belait, at the mouth of the Belait River. An international airport is located at Bandar Seri Begawan.

Administration and social conditions. *Government.* In 1959 Brunei became a self-governing state and adopted a constitution, although the British retained jurisdiction over foreign policy, defense, and internal security. Limited attempts at elected, representative government under this constitution were abandoned by 1970. After Brunei attained full independence in 1984, an Islāmic sultanate was established.

Ultimate authority rests with the sultan, who, as prime minister, presides over a Council of Ministers (cabinet) and is advised by several other councils (Religious, Privy, Succession, and Legislative); the members of these bodies are appointed by the sultan. Judicial power is vested in a Supreme Court, composed of a Court of Appeal and a High Court. There also are religious courts that can appeal to the Religious Council. Brunei is divided into four districts for local administration: Belait, Tutong, and Brunei and Muara in the western enclave and Temburong in the east.

Armed forces. The small, well-equipped Royal Brunei Armed Forces consists mainly of an army group, with smaller navy and air force units. These forces are supplemented by a battalion of British Army Gurkhas and by the Royal Brunei Malay Reserve Regiment, formed in 1988.

Health and welfare. Brunei is a welfare state, with well-developed social facilities. Citizens receive free medical service and education. Medical care in rural areas includes a "flying-doctor" service to the villages, outdoor clinics, and traveling dispensaries, while the capital has a large, modern hospital. The literacy rate has improved considerably since the 1960s, with good educational facilities found throughout the state. The University of Brunei Darussalam was founded in 1985.

Education

For statistical data on the land and people of Brunei, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Brunei's early history is obscure, but it was known to be trading with and paying tribute to China in the 6th century AD. It then came under Hindu influence for a time through allegiance to the Majapahit kingdom in Java. When the ships of Ferdinand Magellan's expedition anchored off Brunei in 1521, the fifth sultan, the great Bolkiah, controlled practically the whole of Borneo, the Sulu Archipelago, and neighbouring islands. Toward the end of the 16th century, Brunei was torn by internal strife. A gradual decline in power continued through the 19th century, notably with the cession of Sarawak to the English adventurer James (later Sir James) Brooke in 1841, the expansion of Sarawak by additional grants to Brooke, the cession to Britain of the island of Labuan in Brunei Bay, and the final loss of what is now Sabah (eastern East Malaysia). Brunei's fortunes began to revive, however, when petroleum was first produced in 1929.

Brunei became a British protectorate in 1888, and in 1906 administration was vested in a British resident, whose advice the sultan was bound to accept. Brunei was occupied (1941–45) by the Japanese during World War II. After the British returned, negotiations began on the eventual independence of Brunei. The first step in this process occurred in 1959, when self-government was achieved and the British resident was replaced by a high commissioner. Britain remained responsible for defense and foreign policy. Brunei adopted a written constitution, and in 1962 a partly elected Legislative Council with limited authority was installed. The conversion to a representative government was interrupted later that year by a revolt, which was suppressed with the help of British forces; the sultan then suspended most provisions of the constitution. New elections were held in 1965, but appointed members still retained their majority in the council.

In 1967 Sultan Omar Ali Saifuddin abdicated in favour of his eldest son, Pengiran Muda Hassanal Bolkiah, although the former sultan continued to exercise influence until his death. Brunei's political life was stable throughout the 1970s in large part because of its flourishing economy and its position as one of the world's wealthiest (on a per capita basis) oil producers. In 1979 the United Kingdom and Brunei signed a treaty whereby Brunei would become fully independent in 1984. Malaysia and Indonesia both gave assurances that they would recognize Brunei's status, thereby allaying the sultan's concern that the state might be incorporated by one of its larger neighbours.

Brunei duly became independent on Jan. 1, 1984, and an Islāmic sultanate was proclaimed. A ministerial form of government was introduced: the sultan became prime minister and held other posts, and he appointed members of his family (including his father as defense minister) to most of the other positions. When his father died in September 1986, the sultan took over the important defense portfolio and enlarged his cabinet. The structure of Brunei's government has been in marked contrast with those of the other ASEAN countries, but Brunei has remained politically stable and economically prosperous.

Independence

(O.J.B.)

For later developments in the history of Brunei, see the BRITANNICA BOOK OF THE YEAR.

Cambodia

Cambodia—officially State of Cambodia (Khmer: Roat Kampuchea)—is a country of Southeast Asia located in the southwestern part of the Indochinese Peninsula. Covering a land area of 70,238 square miles (181,916 square kilometres), it is bordered on the west and northwest by Thailand, on the northeast by Laos, on the east and southeast by Vietnam, and on the southwest by the Gulf of Thailand. The name Cambodia and the French version, Cambodge, are transliterations of the country's traditional name, pronounced Kampuchea in the Khmer, or Cambodian, language.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Cambodia's maximum extent is about 280 miles (450 kilometres) from north to south and 360 miles from east to west. The central region is a low-lying alluvial plain surrounding the Tonle Sap (Great Lake) and the beginnings of the Mekong River delta. Extending outward from this region are transitional plains, thinly forested and with prevailing elevations no higher than several hundred feet above sea level. On the north, along the border with Thailand, the Cambodian

plain abuts a sandstone escarpment that marks the southern limit of the Dangrek (Khmer: Dângrêk) Mountains. A southward-facing cliff, stretching for more than 200 miles from west to east, rises abruptly from the plain to heights ranging from 600 to 1,800 feet (180 to 550 metres), forming a natural frontier boundary. East of the Mekong the transitional plains gradually merge with the eastern highlands, a region of forested mountains and high plateaus that extend into Laos and Vietnam. In southwestern Cambodia two distinct upland blocks, comprising the Krâvanh (Cardamom) Mountains and the Dâmrei (Elephant) Mountains, form another highland region that covers much of the land area between the Tonle Sap and the Gulf of Thailand. In this remote and largely uninhabited area is found Mount Aôral (5,949 feet; 1,813 metres), Cambodia's highest peak. The southern coastal region adjoining the Gulf of Thailand is a narrow lowland strip, heavily wooded and sparsely populated, which is isolated from the central plain by the southwestern highlands.

Drainage and soils. The two most dominant topographical features of Cambodia are the Mekong River and the Tonle Sap. Rising in the Plateau of Tibet and emptying into the South China Sea, the Mekong enters Cambodia from Laos at the Khone Falls and flows generally southward to the border with Vietnam, a distance inside Cambodia of approximately 315 miles. The Tonle Sap, joined to the Mekong by the Sab River, serves as a reservoir for the Mekong. During the rainy season (mid-May to early October), the Mekong's enormous volume of water backs up the connecting river for a distance of 65 miles and flows into the Tonle Sap, expanding the lake's surface area from the dry-season minimum of 1,200 square miles to a rainy-season maximum of more than 3,000 square miles. As the water level of the Mekong falls during the dry season, the process is reversed. Water drains from the Tonle Sap back into the Mekong, reversing its directional flow. As a result of this annual phenomenon, the Tonle Sap is one of the world's richest sources of freshwater fish.

Tonle Sap

Most of Cambodia's soils are sandy and relatively poor in nutrients. The so-called red-soil areas in the eastern part of the country, however, are suitable for such commercial crops as rubber and cotton. The annual flooding of the Mekong during the rainy season deposits a rich alluvial sediment that accounts for the fertility of the central plain.

Climate. Cambodia's climate is governed by the monsoon winds, which define two major seasons. From mid-May to early October, the strong prevailing winds of the southwest monsoon bring heavy rains and high humidity. From early November to mid-March, the lighter and drier winds of the northeast monsoon bring variable cloudiness, infrequent precipitation, and lower humidity. The weather between these seasons is transitional. Maximum temperatures are high throughout the year, ranging from about 82° to 83° F (28° C) in January, the coolest month, to about 95° F (35° C) in April. Annual rainfall varies considerably throughout the country, from more than 200 inches (5,000 millimetres) on the seaward slopes of the southwestern highlands to about 50–55 inches in the central lowland region. Three-fourths of the annual rainfall occurs during the months of the southwest monsoon.

Plant and animal life. Much of Cambodia is heavily forested. The central lowland region is covered with rice paddies, fields of such dry crops as corn (maize) or tobacco, tracts of tall grass and reeds, and thinly wooded areas. Savanna (grassy parkland) is predominant on the transitional plains, with the grasses reaching a height of five feet. In the eastern highlands the high plateaus are covered with deciduous forest and grassland. Broad-leaved evergreen forests grow in the mountainous areas to the north, with trees 100 feet high emerging from thick undergrowths of vines, rattans, palms, bamboos, and assorted woody and herbaceous ground plants. In the southwestern highlands, open forests of pines are found at the higher elevations, while the rain-drenched seaward slopes are blanketed with virgin rain forest growing to heights of 150 feet or more. Vegetation along the coastal strip ranges from evergreen forests to nearly impenetrable mangrove forests.

About three-fourths of Cambodia was forested in 1970. This area has been reduced considerably since then, especially in the provinces bordering Thailand and Vietnam that have been heavily exploited by the local inhabitants and by Thai and Vietnamese entrepreneurs. Unchecked deforestation in northern areas near the Mekong has produced extensive erosion along the river's banks.

The northeastern forests of Cambodia—like the neighbouring areas of Laos and Vietnam—once sheltered large populations of such wild animals as elephants, wild oxen, rhinoceroses, and several species of deer, but the loss of forest cover, combined with warfare and unregulated hunting in the region, has sharply reduced these numbers. Small populations of most species still may be found, along with some tigers, leopards, bears, and many small mammals. Among the more common birds are herons, cranes, grouse, pheasant, peafowl, pelicans, cormorants, egrets, and wild ducks. Four varieties of snakes are especially dangerous: the Indian cobra, the king cobra, the banded krait, and Russell's viper.

Reduction of wildlife

Settlement patterns. Cambodia always has been overwhelmingly a land of villages. Only a small fraction of the total population has ever lived in towns of more than 10,000 inhabitants. Since the 1920s, most of these urban dwellers have been concentrated in Phnom Penh, which is situated at the confluence of the Mekong, Basăk (Bassac), and Sab rivers.

Until the mid-1970s, the vast majority of Cambodia's people inhabited the central lowland region, where the rural village was second only to the family as the basic social unit. The typical Cambodian village in those days was made up of ethnically homogeneous people and had a population of fewer than 300 persons. The village (*phum*) was part of a hamlet or community (*khum*) with which it shared one or more Buddhist temples (*wat*), an elementary school, and several small shops. Cambodian villages usually developed in a linear pattern along waterways and roads, but often houses also were dispersed through the countryside on largely self-contained paddy farms. Houses in Cambodia generally were built on wooden pilings and had thatched roofs, walls of palm matting, and floors of woven bamboo strips resting on bamboo joists. More prosperous houses, while still on pilings, were built of wood and had tile or metal roofs.

Village life

There were a few large landowners in Cambodia, until, under the rulers of Democratic Kampuchea, they were forced off their land and into collectives in 1975 and made to live as ordinary peasants; hardly any reemerged after decollectivization. Before collectivization, the typical villager owned and worked enough land to provide for his family, with a small surplus that could be converted into cash for additional goods or the payment of taxes. Landholdings tended to be small in the crowded south-central regions of the country. During the 1960s the government

M.P. Kahl/Bruce Coleman Ltd

A traditional rural settlement on the bank of the Tonle Sap, Cambodia.

of Prince Norodom Sihanouk was successful in colonizing frontier regions, especially in the northwest, with army veterans or poor farmers from more crowded parts of the country. These programs, however, did not significantly alter Cambodian settlement patterns.

Throughout rural Cambodia, lifestyle was closely geared to the agricultural cycle, which was based, in large part, on family-oriented subsistence farming. Family members were awake before dawn, and most of the day's work was accomplished before noon, although minor tasks were performed in the cool of the early evening. Electricity has always been rare in village areas, and country people were generally asleep soon after sunset. During the rice-growing season, all family members worked together in the fields, because the work of planting, transplanting, and harvesting had to be done quickly. Without mechanical assistance, the work of several people was needed to grow enough rice to feed a family for a year. Because of the intensive labour requirements of paddy farming, obligations would build up among families within a village during the agricultural season. Festivals and marriages, celebrated by a whole village, were usually held after the rice had been harvested and money had been obtained from selling the surplus grain.

Urban areas

The urban areas of Cambodia took their present form in the early 20th century, during the French colonial period, as commercial and administrative centres serving their surrounding rural regions. Most of them were located at the intersections of land or river routes and were relatively accessible to the areas they served. Phnom Penh (*phnom* means "hill"; Penh is a woman's name) is Cambodia's single metropolis, and its population fluctuations reflect the country's recent history. Before the outbreak of war in 1970, it held about 500,000 people; its population by 1975, then swollen with refugees, numbered some 2,000,000. People began moving back into the city in 1979, after its official depopulation. Phnom Penh's population has grown rapidly since then, exceeding its 1970 level in the late 1980s.

The people. Cambodia's first national census as an independent nation was taken in 1962, with a resulting total of about 5,700,000. Overall population density has been difficult to determine, because of the enormous losses and movements of people in the years after 1970. Because so much of the country is poorly watered and without inhabitants, the actual density in populated areas is quite high.

Ethnic and linguistic composition. The Khmer (Cambodians) account for the vast majority of the total population. This has produced a homogeneity that is unique in Southeast Asia and has encouraged a strong sense of national identity. Other traditional ethnic groups included the Chinese, Vietnamese, Cham-Malays, various tribal peoples, and Europeans. With the upheavals of the 1970s, the number of European residents declined precipitously, while many Chinese and local Vietnamese survivors emigrated overseas.

The Khmer are concentrated in the lowland regions surrounding the Mekong River and the Tonle Sap, on the transitional plain, and along the coast. They belong to the Mon-Khmer ethnolinguistic group. A product of centuries of intricate cultural and racial blending, the Khmer moved southward before 200 BC into the fertile Mekong delta from the Khorat Plateau of what is now Thailand. They were Indianized by successive waves of Indian influence and in the 8th century AD were exposed to Indo-Malayan influences and perhaps immigration from Java. This was followed by migrations of Tai peoples from the 10th to the 15th century, by a Vietnamese migration beginning in the 17th century, and by Chinese migrations in the 18th and 19th centuries. The typical Khmer family before 1975 consisted of a married couple and their unmarried children. Both sons and daughters usually left the parental home after marriage to establish their own households.

The ethnic minorities

Among the ethnic minorities in Cambodia before 1975, the Chinese were the most important, for they controlled the country's economic life. They were shunted aside in the communist-led revolution of the 1970s and made to become ordinary peasants. Those who did not seek refuge abroad after 1975 and others who subsequently returned regained some of their former influence as urban centres were revived. The Vietnamese minority occupied a somewhat lower status than the Chinese, and most of them fled or were repatriated to Vietnam after 1970. In the 1980s, however, a large number of Vietnamese migrants, many of them former residents of Cambodia, settled in the country. Centuries of mutual dislike and distrust have clouded Vietnamese-Khmer relations, and intermarriage has been infrequent. The next most important minority, the Cham-Malay group known in Cambodia as Khmer Islām, also maintained a high degree of ethnic homogeneity and was discriminated against under the regime of Democratic Kampuchea. Finally, the tribal people of Cambodia, living originally in the forested northeastern part of the country, received slightly better treatment than the Khmer Islām during that period.

Religions. Most ethnic Khmer are Theravāda (Hīnayāna) Buddhists (*i.e.*, belonging to the older and more traditional of the two great schools of Buddhism, the later school being called Mahāyāna). Until 1975 Buddhism was officially recognized as the state religion of Cambodia. Although the social and psychological characteristics often ascribed to the Khmer—individualism, conservatism, patience, gentleness, and lack of concern for material wealth—were often in the eyes of the beholder, they represented Buddhist ideals toward which a large number of Cambodians, especially in rural areas, have continued to aspire. Buddhist precepts, however, do not permeate Cambodian education and ideology as strongly as they did before 1975.

Minority populations were not Theravāda Buddhists. Tribal people were animists, and the ethnic Vietnamese and Chinese were eclectic, following Mahāyāna Buddhism, Taoism, and such syncretic Vietnamese religious movements as the Cao Dai. The Cham were strict Muslims, and a sizable number of Vietnamese were members of the Roman Catholic church.

Demographic trends. In common with many developing countries, Cambodia has a large proportion of children in its population. It should be noted, however, that population trends have been difficult to trace because of the destruction and dislocation since the early 1970s, because of the lack of statistical information available, and because an enormous number of people of childbearing age have died.

The war and social revolution in the 1970s and the political and economic disruption in the country since then have seriously affected the distribution of Cambodia's population. Between 1975 and 1978, hundreds of thousands of urban people were forcibly moved into rural areas to cultivate rice and to dig and maintain extensive irrigation works. With the exception of Phnom Penh and Bătdâmbâng (Battambang), few towns and cities subsequently have regained their pre-1970 population levels. In addition, more than 600,000 people have sought permanent or temporary residence outside Cambodia since 1979. About a third of these have emigrated overseas, while most of the remainder have settled in refugee camps in Thailand along the Cambodian border; the repatriation of these refugees under United Nations auspices was one of the provisions of a peace agreement signed in Paris in 1991.

Emigration

The economy. Even before 1975, Cambodia's economy was one of the least-developed of the Southeast Asian region. It was heavily dependent on two major products—rice and rubber—and consequently was vulnerable to profound annual fluctuations caused by vagaries in the weather and world market prices. Agriculture dominated the economy, with most rural families engaged in rice cultivation. Although the tradition of landownership was strong, family landholdings were relatively small. But, even with small family farms, the rural population was largely self-sufficient. Two and a half acres (one hectare) of rice paddy provided for the needs of a family of five people, and supplementary requirements were traditionally satisfied by fishing, cultivating fruit and vegetables, and raising livestock. Famine was rare in Cambodia, but the self-sufficiency of the rural family produced a conservatism that proved resistant to government efforts before 1975 to modernize the country's primitive agricultural methods.

The collectivization of agriculture under the rulers of Democratic Kampuchea was dismantled after the collapse of that regime, but it remained an ideal of the communist-led government that came to power in 1979. Voluntary cooperative groupings called *krom samaki* subsequently replaced collective farms in many areas, but the vast majority of Cambodian farming continued to be carried out by family units growing crops for subsistence and small surpluses for cash or barter. A law enacted in 1989 permitted Cambodians to buy and sell real estate for the first time. An immediate effect of the law was a speculative boom in urban areas and an increase in investment, particularly in Phnom Penh. In rural areas laws also were implemented that restored traditional rights of land tenure and inheritance.

Resources. Cambodia has few known mineral resources. Some limestone and phosphate deposits are found in Kâmpôt province, and precious stones are mined in Bătdâmbâng province. Cambodia's small quantities of iron and coal have not justified commercial exploitation. Electric power sources are mainly dependent on imported oil. Prospecting for oil and natural gas has been initiated at offshore areas adjacent to sites being exploited by Vietnam.

Agriculture and fisheries. Rice is Cambodia's principal food, its major crop, and, in times of peace, its most important export commodity. Rice is grown on most of the country's total cultivated land area. The principal rice regions surround the Mekong and the Tonle Sap, with cultivation particularly intensive in Bătdâmbâng, Kâmpóng Cham, Takêv, and Prey Vêng provinces. Cambodia traditionally has produced only one rice crop per year because it has lacked the extensive irrigation system needed for double cropping.

Rice production

Under the government of Democratic Kampuchea, great effort was made to build irrigation systems throughout the country. The results occasionally were notable, and in a few parts of the country farmers were able to grow two, and more rarely three, crops of rice per year. In some cases the works were poorly conceived and hastily built and soon collapsed. Most of those that survived were abandoned after 1979.

Under the traditional patterns of agriculture, planting normally begins in July or August, and the harvest period extends from November to January. The amount of rainfall, when there is little irrigation, determines the size and quality of the crop. Other food products include corn (maize), beans, soybeans, and sweet potatoes. The principal fruit crops, all of which are consumed locally, include oranges, bananas, and pineapple; these are supplemented by a variety of other tropical fruits, including breadfruit, mango, mangosteen, and papaya.

Fisheries and livestock are important components of the domestic economy. Fish in its various forms—fresh, dried, smoked, and salted—constitutes the most important source of protein in the Cambodian diet, and subsistence fishing is part of every farmer's activity. The annual freshwater catch includes perch, carp, lungfish, and smelts. Cattle, particularly water buffalo, are used principally as draft animals in the rice paddies and fields. Hog production has also played a large role in agriculture. Efforts to replenish the number of livestock—depleted by years of war—have been hampered by uncertain social conditions and the prevalence of animal diseases.

Industry. Although industrial development remains at a low level, successive governments since independence have made efforts to build a modest industrial base suitable for domestic needs. Equipment installed in the 1950s and '60s in many Cambodian factories has become obsolete, however, and the country has received little foreign investment capital needed for long-term industrial development. Cambodia also has lacked a trained and experienced labour force. Some progress has been made—plants have been established to produce soft drinks, paper, cigarettes, building materials, cement, and cotton textiles—but Cambodia's industrial sector has found it difficult to compete with mass-produced goods from the more economically developed countries of the region. Timber processing and rice milling, which were important before 1975, were revived in the 1980s.

Finance and trade. Foreign trade has been at a standstill since the 1970s, and imports have taken the form of foreign aid. Traditionally, the bulk of Cambodia's exports—consisting almost entirely of rice, rubber, corn, and other agricultural products—went to other Asian nations, while imports came mainly from Japan, the United States, and western Europe.

The civil war of 1970–75 in Cambodia devastated the countryside, sharply reduced foreign trade, and destroyed the country's fragile economic infrastructure. After the communist victory in 1975, the economic policies of the Democratic Kampuchea regime met with mixed results and left the people with few resources with which to fight a full-scale war against Vietnam in 1978–79. The port of Kâmpóng Saôm (formerly Sihanoukville) was used by Democratic Kampuchea and subsequent regimes largely to receive foreign aid. Substantial maritime trade also has developed between Thailand and Cambodia in the Cambodian province of Kaôh Kŏng.

Effects of the civil war

Transportation. Cambodia's inland waterways and road systems constitute the main transportation routes. Each invariably is affected by the floods of the rainy season, which result in heavy accumulations of silt and washouts caused by flash floods. Railroads rank third in significance. Domestic shipping and civil air facilities are limited. Maritime commerce is carried out almost exclusively by foreign vessels.

The road system eventually surpassed the country's inland waterways as the principal means for moving cargo and passengers. The network was originally designed by the French during the protectorate period to link the agricultural hinterland with the port of Saigon (now Ho Chi Minh City, Vietnam). Consequently, the system did not serve Cambodia as a whole. Extensive land tracts in the northern, northeastern, and southwestern parts of the country were without roads. Of the total road network, only a fraction has been paved; other roads have been surfaced with crushed stone, gravel, or laterite or have been simply graded without being paved. The country's longest bridge, traversing the Sab River at Phnom Penh, was destroyed in 1975.

Roads and bridges deteriorated sharply during the Democratic Kampuchea period and in the civil war that followed. Funds and equipment for repairs were not available, and after 1979 most roads were mined or cut by guerrillas hostile to the government in Phnom Penh. Since 1990, repairing Cambodia's road network has been a high priority for the United Nations and has been a focus of foreign-aid efforts in the country.

Cambodia's inland waterways have a collective extent of some 1,200 miles, of which more than 90 percent are part of the Mekong and Tonle Sap systems. Phnom Penh, located about 200 miles from the mouth of the Mekong, can be reached by oceangoing vessels with drafts of less than 13 feet. North of Phnom Penh, the Mekong is navigable to Krâchéh for rivercraft, but rapids and winding channels in the section between Krâchéh and the Laos border generally preclude commercial navigation.

Cambodia's single maritime port is located at Kâmpóng Saôm on the Gulf of Thailand. Completed in 1960, Kâmpóng Saôm can provide unlimited anchorage for oceangoing ships. The port is of strategic importance to Cambodia, and considerable industrial development has taken place in the area. A modern four-lane highway links Kâmpóng Saôm with Phnom Penh.

The railroad system is owned and operated by the Cambodian government. One line, completed prior to World War II, connects Phnom Penh with Paôy Pêt on the Thai frontier and facilitates the movement of milled rice from the western provinces of Bătdâmbâng, Poŭthĭsăt, and Kâmpóng Chhnăng. Another line, completed in 1969, connects Phnom Penh with Kâmpóng Saôm.

Administration and social conditions. *Government.* In 1981 a constitution was promulgated by the Vietnamese-backed government in Phnom Penh controlled by the communist Kampuchean People's Revolutionary Party (transformed in 1992 into the noncommunist Cambodian People's Party). This document (amended in 1989) provided for a legislative National Assembly and a Council

Constitution

of State selected from the assembly, an executive Council of Ministers, and a judiciary. This government was opposed by three major factions: Prince Norodom Sihanouk and his followers, the communist Party of Democratic Kampuchea (commonly called the Khmer Rouge), and the noncommunist Khmer People's National Liberation Front (renamed the Buddhist Liberal Democratic Party in 1992). In 1982 the three opposition groups formed a coalition government-in-exile headed by Sihanouk, but this rival government never gained power in Cambodia and often was racked by internal dissention.

This arrangement, however, accorded the opposition groups international recognition and gave them the opportunity to engage in discussions with the Phnom Penh government. After lengthy negotiations, a series of peace accords were signed in Paris in 1991 by the four political factions. The agreements provided for the reorganization of the national government under the executive control of a Supreme National Council chaired by Sihanouk and composed of members of the four political factions. A United Nations Transitional Authority in Cambodia, established to oversee the implementation of the accords, was given the administration of several key ministries—including those for defense, information, and the interior. Elections for a new government were held in May 1993.

Armed forces. Each of the four political factions has maintained its own military units, with those of the government in Phnom Penh being the largest and best-equipped. Under the 1991 accords, all factions were to reduce by 70 percent the forces under their control.

Education. Cambodia's educational system, as it had developed in the first 70 years of the 20th century, was another casualty of warfare and ideology. During the Democratic Kampuchea period, only primary schools were open; older students attended irregularly scheduled political and technical courses, often held in the communes. After 1979 the government in Phnom Penh gave high priority to primary education, and it also reopened secondary schools and institutions of higher education. Although a large number of young Khmer attend some form of educational institution, schools and colleges have been severely hampered by shortages of funds, books, equipment, and trained staff.

Health and welfare. Throughout Cambodia's history, an acute shortage of medical personnel has been a major obstacle to the implementation of an effective public health program. Even before the civil war, Cambodia had few doctors, hospitals, or medical facilities. The civil war of 1970–75 strained and eroded this fragile structure. The rulers of Democratic Kampuchea moved medical personnel onto collective farms and, as part of its policy of self-reliance, encouraged non-Western medical practices based on the use of local herbs. In the period since 1979, the scarcity of funds, unsettled conditions in the country, a lack of sanitation, and a shortage of medicine have contributed to reports of malaria and hepatitis epidemics. Phnom Penh has the country's best health care facilities and trained medical personnel; most rural areas are served only by a local infirmary.

Cultural life. Before 1970, Cambodian culture and artistic expression were overshadowed by the greatness of the past. Although the Khmer empire owed much to Indian influence, its achievements represented original contributions to Asian civilization. The magnificent architecture and sculpture of the Angkor period (802–1432), as seen in the temple complexes at Angkor Wat and Angkor Thom, marked the apex of Khmer creativity. Following the capture of Angkor by the Thai (15th century) and the crumbling of the empire, the region underwent four centuries of foreign invasions, civil war, and widespread depopulation. It was not until the establishment of the French protectorate in 1863 that internal security was restored, the country's borders were stabilized, and efforts were undertaken to revive traditional Khmer art forms. After Cambodia gained independence from France in 1953, the government placed particular emphasis on accelerating that revival by establishing a national school of music, a national school of ballet and theatre, and a fine arts university. This coincided with the rapid expansion of elementary and secondary school facilities and the emergence of education as the most important factor of social mobility.

Khmer architecture

While Democratic Kampuchea's leadership, inspired by the People's Republic of China, made culture subservient to Marxist-Leninist doctrines, the government in Phnom Penh after 1979 made serious efforts to restore such traditional activities as classical music, ballet, and popular theatre. Foreign aid from India and Poland was used to clean and maintain some of the temples at Angkor, which had suffered from years of vandalism and neglect. These aspects of high culture have had to compete for people's attention with videotapes imported from Hong Kong, Thailand, and elsewhere, and with Western popular music.

Music and dance forms. Music occupied a dominant place in traditional Cambodian culture. It was sung and played everywhere—by children at play, by adults at work, by young men and women while courting—and invariably was part of the many celebrations and festivals that took place throughout the year at Buddhist temples in the rural countryside. Instruments used in full orchestras included xylophones with wooden or metal bars, one- and two-stringed violins, wooden flutes, oboes, and drums of different sizes. The players followed the lead of one instrument, usually the xylophone, and improvised as they wished.

Dancing and drama were also popular forms of artistic expression. The Royal Ballet in Phnom Penh exemplified the classic, highly stylized dance form adapted by the Khmer and Thai from the ancient dances of Angkor. In the countryside, folk drama and folk dances were performed at festivals and weddings by wandering troupes.

Visual arts. The traditional visual arts of Cambodia revealed the essential conservatism of the Khmer. Ancient themes were preferred and rarely was there an effort to improve or adapt. The principal crafts were weaving, working silver and gold, making jewelry, and the sculpture of wood and stone.

Crafts

Literature. In the 1960s and early 1970s, Cambodia's traditionally conservative literature—which was based to a large extent on Thai literary forms—came under Western influence, as did its audience of young, urbanized Cambodian elite. Novels, poetry, visual arts, and films came to reflect international taste. All these forms of expression, however, were banned by the officials of Democratic Kampuchea, and freedom of expression was limited by the government after 1979 through paper shortages and by the regime's use of literature for propaganda.

Press and broadcasting. No daily newspaper is published in Phnom Penh, but there are several weeklies. Television and radio programs are broadcast for only a few hours per day by government-controlled stations.

(L.C.O./D.P.Ch.)

For statistical data on the land and people of Cambodia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The importance of Cambodia's contributions to mainland Southeast Asia is out of proportion to its present reduced territory and limited political power. Between the 11th and the 13th century, the Khmer (Cambodian) state included much of the Indochinese Peninsula and incorporated large parts of present-day southern Vietnam, Laos, and eastern Thailand. The cultural influence of Cambodia on other countries, particularly Laos and Thailand, has been enormous.

Early history. It is not known for certain how long people have lived in what is now Cambodia, where they came from, or what languages they spoke before writing was introduced (using a Sanskrit-style alphabet) about the 3rd century AD. Dates based on carbon-14 measurements have established that people able to make pottery inhabited Cambodia as early as 4000 BC. These and subsequent finds have suggested that these early people, like Cambodians today, were of slight to medium build, constructed their houses on wooden piles, consumed a considerable quantity of fish, and raised pigs and water buffalo.

Whether the early inhabitants of Cambodia came originally or primarily from the north, west, or south is still

being debated, as are theories about waves of different peoples moving through the region in prehistoric times. The notion of distinct Asian "races" that was popular until the 1950s has been discredited, and subsequent archaeological finds have suggested that prehistoric mainland Southeast Asia, including Cambodia, had a comparatively sophisticated culture. Some scholars even have attributed the first cultivation of rice and the first casting of bronze to the region.

Indian influences

Funan and Chenla. Indian influences were the most important of Cambodia's early history. They coincided with the first centuries AD, when Chinese and Indian pilgrims and traders stopped along the coasts of present-day Cambodia and Vietnam and exchanged silks and metals for spices, aromatic wood, ivory, and gold. Written sources from about this period are almost entirely in Chinese. A kingdom or group of kingdoms known to Chinese writers as "Funan" flourished in southern Cambodia at this time. Over a period of 300 years between the 3rd and 6th centuries AD, its rulers offered gifts from time to time to Chinese emperors. Chinese writers testified to the extent of Indian influence in the kingdom and cited a local story, dating from the 6th century, that traced its origins to an Indian Brahman named Kaundinya "who changed its institutions to follow Indian models." One early innovation was probably the introduction of large-scale irrigation, which allowed people to raise three or more crops of rice per year in some districts and brought unpromising areas under cultivation. Another was the worship of the Hindu god Śiva (Shiva), who was conceptualized as a tutelary ancestor or spirit of the soil and often was represented by a stone lingam, or phallus. A third was the relatively peaceful coexistence of Hinduism and Buddhism, which characterized Cambodia for more than a thousand years.

The most important legacy of Funan, though it may have been exaggerated by Chinese writers, was a relatively centralized state apparatus, culminating in a theoretically absolute ruler. The ruler presided over the agricultural workforce and commanded agricultural surpluses and off-season labour to sustain his lifestyle, support a priestly caste, and build fortresses, palaces, and temples. In a general way, these social arrangements characterized much of the medieval world, but it would be imprecise to use a term like "feudalism" to characterize Funan and its successor states. Instead, it is probably more fruitful to seek links between ancient and present-day Cambodia than between ancient Cambodia and countries far to the west about which the Khmer would have known nothing.

The appearance of Sanskrit inscriptions in the 6th century—the earliest known Khmer inscription dates from the early 7th century—has made it possible to use indigenous sources to supplement Chinese ones, but they all fail to clarify the confusing political developments that occurred in the Cambodian region between the decline of Funan in the 6th century and the founding of a centralized state in northwestern Cambodia about three centuries later. It has been common practice for modern writers to use "Chenla," the contemporary Chinese term for the region, when referring to Cambodia during this time. Chinese sources suggest that there were at least two kingdoms, known as "Water Chenla" and "Land Chenla," in Cambodia that vied for recognition from China in this period. Whereas the geographic centre for both Funan and Water Chenla lay in the Mekong delta south and east of Phnom Penh and extended into present-day Vietnam, the heartland of Land Chenla appears to have been farther up the Mekong, with an important cult site called Wat Phu located in present-day southern Laos. It seems likely that Water Chenla looked outward and welcomed foreign trade, while Land Chenla was more inward-looking and based its economy on intensive agriculture. Surviving inscriptions in Sanskrit and Khmer testify to a multitude of small kingdoms on Cambodian soil in the 7th, 8th, and 9th centuries. Remarkable sculptures and architectural remains also have survived from this period, displaying a mixture of Indian influence and local inspiration. The appearance of local styles reflected, in part, declining Indian commercial interest in the region beginning in the 7th century.

The Khmer state (Angkor). *Foundation of the kingdom.* In 790 a young Cambodian prince, claiming to be descended from the rulers of Funan, was consecrated in eastern Cambodia under the title Jayavarman II. Part of the ceremony involved breaking ties with "Java," which probably was not a reference to the island of Java but to the kingdom of Śrivijaya on the island of Sumatra. Over the next 10 years, Jayavarman extended his power northward into the Mekong River valley until, in 802, he was reconsecrated as a chakravartin (the ancient Indian conception of world ruler) in northwestern Cambodia. The capital seems to have been located in the Kulén Hills north of the present-day provincial capital of Siĕmréab (Siem Reap), where he died in 835. Despite the high status accorded him by subsequent Angkorean kings, Jayavarman II seems to have left no inscriptions of his own, and the monuments that can be dated to his reign were small and hastily built.

Jayavarman II

Jayavarman's real accomplishment was less tangible and longer-lasting, for he seems to have welded what came to be called Kambuja-desa into a confident, self-aware kingdom that superseded and came to control a range of smaller states. He was Cambodia's first nationally oriented king. Whether smaller states were forced into subservience or volunteered it is uncertain, but, despite the grandeur and apparent continuity of the Angkorean temples that were built over the next four centuries, Jayavarman II's successors were often powerless or were constrained by outside forces. Revolts and usurpations were frequent, as were foreign invasions. Rulers were subject to claims by family members, priests, generals, and bureaucrats. Some kings, especially usurpers, had more freedom of action than others. Those who ruled in periods of peace were also in a better position to undertake building programs and public works. Like their counterparts in medieval Europe, Cambodian kings were far removed from ordinary people. The king was perceived primarily in religious terms, and he assured the fertility of the soil and the well-being of the kingdom through the rituals he performed that were thought to obtain such guarantees. In exchange for his protection, the people were technically enslaved and liable for military service and corvée duty, although they actually spent most of their time growing crops to feed their families.

Toward the end of the 9th century, soon after Jayavarman II's death, the Cambodian capital shifted to the northern shores of the Tonle Sap, near present-day Phumĭ Rôluôs. A king named Indravarman I (ruled 877–*c.* 890) constructed a large reservoir and several temples there, including a pyramidical structure called the Bakong—the first Cambodian temple to be built primarily of stone rather than brick. This so-called "temple mountain" became the model for many larger royal temples at Angkor. These served as monuments to the greatness of their patrons and, subsequently, as their tombs.

Angkorean civilization. Indravarman's son and successor, Yaśovarman I (*c.* 890–*c.* 910), moved the capital again, this time closer to Siĕmréab. This was the foundation of Angkor—a name derived from the Sanskrit word *nagara,* meaning "city"—which has become one of the world's most celebrated archaeological sites, as well as the popular name for Cambodia's medieval civilization. The city that Yaśovarman founded, Yaśodharapura, retained that name and remained Cambodia's capital until it was abandoned in the 15th century. His temple mountain, now called Bakheng (literally, "Mighty Ancestor"), was built on a natural hill that overlooked a teeming city, the more distant rice-growing plain, and the Tonle Sap. The mountain occupied the centre of the city, just as Mount Meru, the mythical home in India of Hindu gods, was said to stand at the centre of the universe. Yaśovarman built a large reservoir nearby. The city wall of Yaśodharapura measured 2.5 miles (4 kilometres) on each side. For such an ambitious building program, the king needed to command a large labour pool. Other evidence suggests that his reign was characterized by tolerance toward a variety of Buddhist and Hindu sects that occasionally blended into local cults honouring ancestral spirits and spirits of the soil. Indeed, for all the apparent absolutism of its kings,

Yaśovarman I

a consistent feature of Angkorean civilization, unmatched in medieval Europe, was religious toleration.

After several decades of warfare, dislocations, and disorder—Yaśodharapura itself was abandoned for nearly 30 years—Rajendravarman II (ruled 944–968) restored the capital and set in motion a period of peace and prosperity that lasted nearly a century. During the reign of his successor, Jayavarman V (968–*c.* 1000), the rose-coloured sandstone shrine of Banteai Srei—arguably the loveliest temple at Angkor—was built on the outskirts of the capital under the patronage of a wealthy priestly family, one of whom had been Jayavarman's teacher. In Yaśodharapura itself, Jayavarman V began work on the imposing temple mountain now called Ta Keo, which was completed under his successor, Suryavarman I (ruled *c.* 1004–*c.*1050). Suryavarman I, an innovative and demanding monarch, was a usurper with links to princely families in what is now northeastern Thailand. His rise to power involved the subjugation of many areas that had become semi-independent under his predecessors and resembled that of Jayavarman II two centuries earlier. Suryavarman extended the Khmer empire westward into present-day Thailand, where he constructed the large mountaintop temple known as Preah Vihear. During his reign, the number of cities ruled from Yaśodharapura jumped from roughly 20 to nearly 50, and foreign trade increased, along with tighter bureaucratic control. His successor consolidated these gains, put down a dangerous rebellion, and was responsible for the temple mountain known today as the Baphuon.

The closing years of the 11th century were ones of turmoil and fragmentation. At different times, two and even three "absolute monarchs" contended, simultaneously, for the title of chakravartin. At the end of the century, however, a new dynasty—which was to last for more than a century—began to rule at Angkor. Its most powerful monarch took the name of Suryavarman II (ruled 1113–*c.* 1150), although he probably was not descended from the earlier king of that name. Like his namesake predecessor, Suryavarman II was a formidable military campaigner. He avenged earlier attacks on Angkor by armies launched from the kingdom of Champa, in what is now south-central Vietnam, and led expeditions into northern and southern Thailand. A campaign against Vietnam, which recently had declared its independence from China, was less successful, although earlier in his reign Suryavarman had renewed diplomatic relations with China.

Angkor Wat

Suryavarman's major accomplishment, from a 20th-century perspective, was his temple complex of Angkor Wat, still the largest religious structure in the world and one of the most beautiful. The temple, which eventually became his tomb and probably was an astronomical observatory as well, was dedicated to the Hindu god Vishnu. Its bás-reliefs, running for nearly a half mile inside its third enclosure, depict events in the well-known Indian epics the *Mahābhārata* and *Rāmāyaṇa*—confirming that these texts were widely known at Angkor—as well as Suryavarman himself holding court. The elegance of these carvings, the hundreds of graceful statues of angelic dancers (*apsaras*) that adorn the temple, and its reflection in the moats that surround it continue to give Angkor Wat an awe-inspiring air; in the 12th century, when its towers were gilded and its moats properly maintained, it must have been even more breathtaking.

Jayavarman VII. Suryavarman II's successor, Yaśovarman II, also reached into earlier history for his royal name, tracing his lineage to the Rôluŏs period. During his reign, several temples begun under Suryavarman were completed. Yaśovarman was overthrown by one of his officials after returning from a military campaign in Thailand. In the aftermath of the coup, a Cambodian prince, later to rule under the name of Jayavarman VII (1181–*c.* 1220), hurried home from Champa—it is uncertain from his inscriptions why he was there—to vie for the Cambodian throne. He arrived too late, and for the next 10 years he bided his time as the usurper lost control and Angkor was invaded and occupied by the Chams. In 1177, heading an army of his own, the prince attacked Angkor and defeated the Cham forces. The battles are vividly depicted in the bas-reliefs of his temple mountain, the Bayon. To forestall further Cham attacks, Jayavarman annexed the Cham capital, and Angkor controlled Champa until Jayavarman's death.

When his campaign against the Chams was over, the future monarch worked to bring Cambodia under his control. An inscription referred to the kingdom as being "shaded by many parasols," a metaphor for a multiplicity of rulers. In 1191, presumably when the process was complete, Jayavarman finally settled in Angkor. He soon embarked on a program of building and public works that was more extensive and grandiose than any in Angkorean history. According to his inscriptions, hundreds of thousands of people were involved in these projects.

Khmer empire *c.* 1200.

Numerous temples, statues, stone bridges, and inscriptions in the Angkor region and elsewhere in Cambodia testify to the vigour of Jayavarman VII's long reign. He rebuilt and refortified the city. He was a fervent Buddhist of the Mahāyāna (Greater Vehicle) school, but, like most other Cambodian kings, he also tolerated and patronized Hinduism and local ancestor cults; several larger-than-life-size statues of the monarch depict him in meditation.

The Bayon

His extraordinary temple, the Bayon, with its multiple towers, each bearing faces of divinities turned in the cardinal directions, is perhaps the most intriguing of the monuments at Angkor. Like Yaśovarman I's Bakheng, the Bayon stood at the centre of the royal city—which had shifted since Yaśovarman's time—and symbolized Mount Meru. Many Hindu gods and the Buddha are depicted in the statuary of the temple, while the bas-reliefs depict scenes of ordinary life, providing a picture of 12th-century Cambodians at work, rest, and play that fails to emerge from the religiously oriented inscriptions or from carvings at other temples. The clothing, tools, houses, and oxcarts in the bas-reliefs closely resemble those in the Cambodian countryside today.

The decline of Angkor. After Jayavarman's death about 1220, few monuments were erected at Angkor, and fewer inscriptions were incised. Little by little, the Khmer empire began to contract. Jayavarman's campaigns neutralized Champa as a threat to Angkor, but, by the early 13th century, vigorous new kingdoms in what is now northern Thailand—centring on the city of Sukhothai—became powerful enough to throw off Angkorean domination, as did some Tai principalities in the south. In the mid-13th century, Tai armies even raided Angkor. For most of the

century, however, Angkor remained a glittering, crowded, and wealthy city. It impressed a Chinese visitor, Chou Ta-kuan, who arrived there in 1296. Chou's account is the longest and most detailed description that has survived of the Khmer capital, supplementing the bas-reliefs of the Bayon. He left a picture of a bustling city in which the king still went forth in great pomp and ceremony.

Chou also observed monks of the Theravāda school of Buddhism at Angkor. This more orthodox and austere school flourished in kingdoms to the west of Cambodia and contrasted sharply with the lavish and elitist rituals associated with Hinduism and Mahāyāna Buddhism. When Chou visited Angkor, Theravāda Buddhism was still one religion among many. Soon afterward, however, it began to benefit from royal patronage, and the conversion of the majority of the population probably followed the conversion of members of the elite. Those disadvantaged by the change included the high-ranking priestly families who had built and maintained the temples and had supported the labourers at Angkor.

Some historians have considered the mass conversion to Theravāda Buddhism as having been responsible for the abandonment of Angkor, which certainly accompanied the conversion in the 14th and 15th centuries. That argument has been undermined by the fact that Theravāda Buddhists from Thailand profited from and even accelerated the collapse by their repeated military attacks, in which hundreds and perhaps thousands of Cambodians were led to captivity in Thailand.

Recorded Tai attacks on Angkor occurred in 1369, 1389, and 1431, after which the Khmer capital was definitely abandoned. There undoubtedly were other attacks as well. In 1351 a Tai kingdom whose court modeled itself culturally on Angkor was founded at Ayutthaya (Ayudhya, or Siam), not far from modern Bangkok. The Tai capital remained at Ayutthaya for the next 400 years. It is tempting to imagine a transfusion of elite culture from Angkor to the more prosperous, more secure Tai court in the 14th and 15th centuries. It is also likely that the Khmer who remained at Angkor were drawn southward to the vicinity of Phnom Penh (said to have been founded in the mid-15th century) by the region's commercial possibilities. In any case, the smaller, outward-looking Khmer kingdom that replaced Angkor in the south earned its wealth primarily from trade rather than from intensive rice cultivation and the mobilization of labour for public works.

Ayutthaya

Tai and Vietnamese hegemony. In the century and a half that followed the abandonment of Angkor, what is known of Khmer history is a confusing mixture of uncertain dates, mythical figures, and complex dynastic rivalries. Cambodian chronicles for this period, composed several centuries afterward, are impossible to verify against inscriptions or other primary sources. Between the mid-14th and the end of the 16th century, all that is known for certain is that Angkor was abandoned, that the Tai court of Ayutthaya absorbed some of its culture and prestige, and that the political centre of Cambodia shifted to the south. Relations between the Tai and the Khmer remained uneasy.

In the late 16th century, a period of Tai weakness following wars with Myanmar (Burma) coincided with a time of Cambodian prosperity; and a Khmer monarch, Chan I (ruled 1516–66), reoccupied the Angkor area briefly, restoring some of the temples, adding some bas-reliefs to those at Angkor Wat, and leaving several new inscriptions. When the Tai recovered their strength in the 1590s, however, they invaded Cambodia in force and sacked the Khmer capital at Lovek, north of Phnom Penh, ushering in a period of Cambodian weakness vis-à-vis its neighbours that has endured to the present day.

Cambodian history from the beginning of the 17th century until the establishment of the French protectorate in 1863 is, indeed, a sorry record of weak kings being undermined by members of their families and forced to seek the protection of their stronger neighbours, Siam (Thailand) and Vietnam. Between 1603 and 1848, 22 monarchs occupied the Cambodian throne. The details of this unstable, humiliating period are perhaps less important than the record that is available of the manner in which Cambodia slowly fell under the suzerainty of its two neighbours. When a Cambodian monarch in the 1620s foolishly declared "independence" from Ayutthaya, for example, he sought assistance from the Nguyen overlords of southern Vietnam. In exchange, he was encouraged to marry a Nguyen princess and also to permit Vietnamese settlers to move into land near present-day Ho Chi Minh City (Saigon), which until then had been under his control. Over the next 200 years, Cambodian kings sought Tai or Vietnamese protection against their rivals in the royal family and against the foreign power temporarily out of favour. The costs in lost territory and diminished autonomy were considerable, but the monarchs had little bargaining power and no freedom to maneuver.

That Cambodia survived at all can be attributed to the fact that in the 18th century the Tai and the Vietnamese had other preoccupations. In the 1750s and '60s, Tai energies were taken up by wars with Myanmar, whose armies sacked and destroyed Ayutthaya in 1767. Soon afterward, the Nguyen rulers of southern Vietnam were engaged in a prolonged campaign to regain power from the usurping Tay Son rebels. Fighting spilled over from Vietnam into Cambodia, and the royal family fled to Thailand. By the end of the century a powerful Tai dynasty had established the kingdom of Siam and had installed itself in its new capital in Bangkok, and at the beginning of the 19th century the Nguyen founded a dynasty that governed all of Vietnam. A confrontation between the two powers in Cambodia was inevitable. In 1794, in exchange for placing a refugee Cambodian prince, Eng, on the Cambodian throne, the Siamese appropriated two Cambodian provinces, Bătdâmbâng (Battambang) and Siĕmréab (Siem Reap)—the latter including the abandoned ruins of Angkor. These provinces remained in Siamese hands until 1907. When Eng died after a short reign, he was replaced by his young son, who ruled as Chan II with Thai protection.

Loss of territory

Chan II's reign confirmed Cambodia's dual vassalage to Thailand and Vietnam. With three rebellious younger brothers and demanding patrons at the Siamese court, he sought assistance from Vietnam; the Siamese supported his brothers, who took refuge in Bangkok. The uneasy calm that ensued, with Chan acknowledging Siamese and Vietnamese suzerainty, ended with Chan's death in 1835. Vietnamese pressure was strong enough to ensure that a powerless princess named Mei was then enthroned, while the Vietnamese controlled most of the country. Not until 1841, when Chan's brother Duong (Duang; ruled 1848–60) returned from exile in Bangkok supported by Siamese troops, were the Cambodians able to exercise a small degree of independence. Fighting between the Siamese and Vietnamese continued in Cambodia for several years. Duong was crowned only after Vietnamese troops agreed to leave the country. Cambodia again became a Siamese protectorate. Duong tried hard to revitalize the kingdom's institutions, but his resources were desperately limited, and his reign was marred by several rebellions. When he died, he was succeeded by his son, Norodom, but conditions were too unstable in the kingdom for Norodom to be crowned.

French rule. *The protectorate.* French control over Cambodia was an offshoot of French involvement in the neighbouring provinces of Vietnam. Their decision to advance into Cambodia came only when they feared that British and Siamese expansion might threaten their access to the largely unmapped Mekong River, which they assumed (incorrectly) would provide them with access to central China. In 1863 French naval officers from Vietnam persuaded Norodom to sign a treaty that gave France control of Cambodia's foreign affairs. The effect of the treaty was to weaken Siamese protection. A French admiral participated in Norodom's coronation, with Siamese acquiescence, in 1864.

For the next 15 years or so, French protection was not especially demanding, and Norodom benefited from French military help in putting down a series of rebellions. By the late 1870s, however, French officials in Cambodia were pressing for greater control over internal affairs. Shocked by what they regarded as the ineptitude and bar-

barity of Norodom's court and anxious to turn a profit in Cambodia, they sought to introduce fiscal and judicial reforms. In doing this, the French knew that Norodom's half brother, Si Votha (Sisowath), who had ambitions for the throne, would cooperate with them. Norodom, however, resisted the reforms, which he correctly perceived as infringements on his power. Exasperated by his intransigence, the French in 1884 forced him at gunpoint to sign a document that virtually transformed Cambodia into a colony. Soon thereafter, provincial officials, feeling threatened, raised guerrilla armies to confront the French.

Anti-French rebellion

The rebellion, which lasted until mid-1886, was the only anti-French movement in the kingdom until after World War II. The French succeeded in suppressing it after agreeing to some concessions to the king, but Norodom's apparent victory was hollow. What the French had been unable to achieve by the convention of 1884, they proceeded to gain through piecemeal action. As Norodom's health declined and as senior Cambodian officials came to see their interests increasingly linked with French power, the way was opened for greater French control. In 1897 the French representative in Phnom Penh assumed executive authority, reducing the king's power to a minimum. Norodom died, embittered and overtaken by events, in 1904.

The first 40 years of the French protectorate—whatever French motives may have been—had guaranteed the survival of the Cambodian state and had saved the kingdom from being divided between its two powerful neighbours. Norodom's successor, Sisowath (ruled 1904–27), was more cooperative with the French and presided benignly over the partial modernization of the kingdom. The northwestern provinces of Bătdâmbâng and Siĕmréab were returned to Cambodia by the Siamese in 1907. By the time Sisowath died 20 years later, hundreds of miles of paved roads had been built, and thousands of acres of rubber plantations had been established by the French. Resistance to their rule, in sharp contrast to what was happening in neighbouring Vietnam, was almost nonexistent.

Sisowath's eldest son, Monivong, who reigned until 1941, was even more of a figurehead than his father had been. During the 1930s, a railway opened between Phnom Penh and the Siamese (Thai) border, while the first Cambodian-language newspaper, *Nagara Vatta* ("Angkor Wat"), affiliated with the Buddhist Institute in Phnom Penh, conveyed a mildly nationalist message to its readers.

World War II and the First Indochina War. When Monivong died in 1941, Japanese forces already had occupied the component states of the Indochinese Peninsula, while leaving the French in nominal control. In these difficult circumstances, the French governor-general, Jean Decoux, placed Monivong's grandson, Prince Norodom Sihanouk, on the Cambodian throne. Decoux was guided by the expectation that Sihanouk, then only 18 years old, could be easily controlled. In the long run, the French underestimated Sihanouk's political skills, but for the remainder of World War II he was a pliable instrument in their hands.

The effect of the Japanese occupation on Cambodia was less profound than it was elsewhere in Southeast Asia, but the overthrow of the French administration by the Japanese in March 1945, when the war was nearing its end, provided Cambodians with opportunities for political development. Pressed by the Japanese to do so, Sihanouk declared his country's independence, and for several months the government was led by Son Ngoc Thanh, a former editor of *Nagara Vatta,* who had been forced into exile in Japan in 1942.

Early postwar period

In October 1945, after the war was over, the French returned to Indochina, arrested Son Ngoc Thanh, and reestablished their control. Cambodia soon became an "autonomous state within the French Union," with its own constitution and a handful of political parties, but real power continued to rest in French hands. Between 1945 and the achievement of complete independence in 1953, however, several significant political developments occurred. The most important was the confrontation between Sihanouk and his advisers and the leaders of the pro-independence Democratic Party, which dominated the National Assembly. Cambodia was poorly prepared for parliamentary democracy, and the French were unwilling to give the National Assembly genuine power. The Democrats, for their part, suffered from internal dissension. The death in 1947 of their leader, Prince Yuthevong, was a severe blow, exacerbated by the assassination of Yuthevong's heir apparent, Ieu Koeuss, in early 1950. Outside Parliament, Son Ngoc Thanh, released from exile in France in 1951, formed a dissident movement, the Khmer Serei ("Free Khmer"), that opposed both Sihanouk and the French.

In June 1952 Sihanouk assumed control of the government. Many Cambodian students in France, among them Saloth Sar (who would become the future communist dictator Pol Pot), objected to Sihanouk's move, but inside Cambodia the king remained extremely popular. His self-styled "Royal Crusade" (a tour of several countries to elicit their support) wrested political independence from the French, anxious to compromise in any case, at the end of 1953. Sihanouk's success discredited the communist-dominated guerrilla movement in Cambodia—associated with the Viet Minh of Vietnam—and Son Ngoc Thanh's anticommunist Khmer Serei.

Independence. At the Geneva Conference convened in 1954 to reach a political settlement to the First Indochina War, Sihanouk's government was recognized as the sole legitimate authority within Cambodia. This decision prevented the Viet Minh from gaining any regional power in Cambodia, as they did in Laos.

While they recognized Sihanouk's role in gaining Cambodia's independence, Democrats and communists alike opposed his increasing authoritarianism. Unable to govern unopposed, Sihanouk abdicated the throne in March 1955 in favour of his father, Norodom Suramarit, and formed a mass political movement, the Sangkum Reastr Niyum ("People's Socialist Community"), whose members were forbidden to belong to other political parties. The effect of the move was to draw thousands of people away from the Democrats, who had expected to win the national elections scheduled for later in the year. When the elections took place, amid widely reported abuses by Sihanouk's police, the Sangkum won every seat in the National Assembly. From then until he was overthrown in 1970, Sihanouk was the central figure in Cambodian politics, sometimes as prime minister and—after his father's death in 1960, when no new monarch was named—as head of state. Overt political life was strictly controlled by the prince, his colleagues, and the police. Cambodian communists, a marginal group of fewer than a thousand members, operated clandestinely and enjoyed little success. In 1963 Saloth Sar, a schoolteacher who also was secretary of the party, fled Phnom Penh and took refuge in the forests along the Vietnamese border; from there he built the organization that later would be known as the Khmer Rouge.

Origin of the Khmer Rouge

Until the mid-1960s, when opposition to his rule intensified, Sihanouk was widely revered in Cambodia. He saw Thailand and what was then South Vietnam as the greatest threats to Cambodia's survival. These two countries were allied with the United States, which the prince distrusted. At the same time, Sihanouk feared the eventual success of the Vietnamese communists in their war against South Vietnam and the United States and was worried by the prospect of a unified Vietnam under communist control. To gain some freedom to maneuver, he proclaimed a policy of neutrality in international affairs. Convinced, however, of American involvement in two South Vietnamese-backed plots against the Cambodian state in 1959 and encouraged in his anti-Americanism by the French president Charles de Gaulle, whom he idolized, Sihanouk broke off relations with the United States in 1965. Soon afterward, he concluded secret agreements with the Vietnamese communists, who were allowed to station troops on Cambodian territory in outlying districts as long as they did not interfere with Cambodian civilians. The secret agreement protected Sihanouk's army from attacks by the Vietnamese but compromised his neutralist policies. After 1965, when the war in Vietnam intensified, he also edged toward an alliance with China.

Cambodia's internal politics after 1965 developed in a complex fashion. Elections in 1966, the first since 1951 not to be stage-managed by the prince, brought in a majority of National Assembly members who owed little or nothing to Sihanouk himself. Although the prince was still a revered figure among the rural populace, he became increasingly unpopular with the educated elite. Conservatives resented his break with the United States and his seemingly procommunist foreign policy, while Cambodian radicals opposed his internal policies, which were economically conservative and intolerant of dissent. A rebellion in Bătdâmbâng province in 1967, manipulated by local communists, convinced the prince that the greatest threat to his regime came from the radical sector, and without hesitation he began using severe measures—including imprisonment without trial, assassinations, and the burning of villages—to impose his will.

By 1969 Sihanouk's grip on Cambodian politics had loosened, and conflict between his army and communist guerrillas, especially in the northeast, had increased. Some anticommunist ministers led by Prince Sirik Matak and General Lon Nol plotted to depose Sihanouk, whose credibility with radicals had evaporated following his renewal of diplomatic relations with the United States. Sihanouk's elaborate policy of juggling major powers against each other had failed. Matak and Lon Nol worked closely with anticommunists in South Vietnam, including Son Ngoc Thanh, whose Khmer Serei movement had gained recruits among the Khmer-speaking minority in Vietnam.

Civil war. In March 1970, while the prince was visiting the Soviet Union, the National Assembly voted to remove him from office as chief of state. Confused and hurt, Sihanouk traveled on to Peking and accepted Chinese advice to resist the coup by taking charge of a united front government-in-exile. This government was to be allied with China and North Vietnam and was to use the Cambodian communist forces led by Saloth Sar, which only a few days before had been fighting against Sihanouk's army.

Lon Nol's government

In Phnom Penh, Lon Nol's new government was initially popular, particularly for his quixotic pledge to rid Cambodia of Vietnamese communist troops. In fact, the confrontation dragged Cambodia fully into the Vietnam conflict. In May 1970 an American and South Vietnamese task force invaded eastern Cambodia, but communist forces already had retreated to the west. Two offensives launched by Lon Nol—named for the semimythical Cambodian kingdom of Chenla—were smashed by the Vietnamese, and thereafter he assumed a defensive stance. North Vietnamese support for the communists diminished in 1973, following the cease-fire agreement reached in Paris with the Americans. The Cambodian communists, however, refused to adhere to the agreements, and in 1973 they were subjected to a massive American aerial bombardment—this occurring despite the fact that the United States and Cambodia were not at war and that no American troops were endangered by Cambodia. The bombing slowed communist attacks on Phnom Penh and wreaked havoc in the heavily populated countryside around the capital. The civil war lasted for two more years, but already by the end of 1973 the Lon Nol government controlled only Phnom Penh, the northwest, and a handful of provincial towns.

In the meantime, Sihanouk declined in importance. By the end of 1973 the Cambodian communists controlled every aspect of the resistance, although they still claimed Sihanouk as a figurehead. Lon Nol's isolated regime in Phnom Penh continued to receive large quantities of American aid, increasing opportunities for corruption.

In April 1975 the Lon Nol government collapsed. Communist forces entered Phnom Penh and immediately ordered its inhabitants to abandon the city and take up life in rural areas. Phnom Penh and other cities and towns throughout the country were emptied in less than a week. Thousands of city dwellers died on the forced marches, but in subsequent years conditions worsened.

Democratic Kampuchea. Over the next six months, following the directives of a still-concealed Communist Party of Kampuchea, Cambodia experienced the most rapid and radical social transformation in its history. Money, markets, and private property were abolished. Schools, hospitals, shops, offices, and monasteries were closed. Nothing was published, no one could travel without permission, and everyone was ordered to wear peasant work clothes. As in Mao Zedong's China, the poorest peasants were favoured at everyone else's expense—an irony that probably escaped the largely bourgeois leaders of the communist movement themselves. A handful of these men and women controlled everything in the country, but they remained in hiding and explained few of their decisions. Instead, they urged everyone to "build and defend" the country. In April 1976 Sihanouk resigned as head of state, soon after a new constitution had renamed the country Democratic Kampuchea. An unknown figure named Pol Pot became prime minister, and more than a year passed before observers outside the country were able to identify him as Saloth Sar.

Khmer Rouge policies

In 1976–77 the new regime, following the lead of Maoist China, sought to achieve the total collectivization of Cambodia, mobilizing its population into an unpaid labour force and seeking to double the average prerevolutionary yields of rice immediately and on a national scale. The human costs of this ill-conceived experiment were enormous. Conservative estimates are that between April 1975 and early 1979, when the regime was overthrown, at least one million Cambodians—about 15 percent of the total population—died from overwork, starvation, disease, or execution. Parallels have been drawn between these events and the fate of European Jews in World War II, Mao's Great Leap Forward in China in the late 1950s, and Josef Stalin's collectivization of Ukrainian agriculture in the Soviet Union in the 1930s. The Soviet and Chinese experiments, rather than the European Holocaust, appear to have been used as models by the Khmer Rouge, although the proportion of the population killed in Cambodia was greater than it had been in China or the Soviet Union. The number of deaths stemmed from the literalism with which plans were carried out, the cruelty of the inexperienced communist cadres, and—as far as executions were concerned—the suspicions of the leadership that the failure of their experiment could be traced to "traitors" in the pay of foreign powers. The Communist Party's interrogation centre in Phnom Penh was the site of more than 20,000 such executions. Those tortured and put to death included many who had served the party faithfully for years—victims of the extreme paranoia of Pol Pot and his colleagues.

Vietnamese intervention. The Khmer Rouge initially had been trained by the Vietnamese, but since the early 1970s they had been resentful and suspicious of Vietnam and Vietnamese intentions. Scattered skirmishes between the two sides in 1975 had escalated into open warfare by the end of 1977. Despite continuing infusions of Chinese aid, the Cambodians were no match for the Vietnamese forces. In December 1978 a large Vietnamese army moved into Cambodia, brushing aside the Democratic Kampuchean forces. Within two weeks the government had fled Phnom Penh for Thailand, and the Vietnamese had installed a puppet regime—the People's Republic of Kampuchea—consisting largely of Cambodian communists who had deserted Pol Pot in 1977–78.

Over the next decade, under the relatively benign tutelage of the Vietnamese, Cambodia struggled back to its feet. Private property was restored, schools and Buddhist practices were reintroduced, cities were repopulated, and, with freedom of movement, trade flourished. At the same time, at least 500,000 Cambodians, some one-fifth of them associated with the communists, fled to Thailand in the aftermath of Democratic Kampuchea's fall and because of the hardship, uncertainty, and disorder that accompanied the installation of the new regime. Of these, perhaps 200,000 people, including most of the surviving members of Cambodia's educated elite, sought refuge in other countries, while the rest came under the control of three resistance groups camped along the Thai-Cambodian border: Norodom Sihanouk and his followers, the Khmer Rouge, and the noncommunists under the leadership of Son Sann (a former prime minister). These groups were supported financially by foreign powers anxious to oppose

Vietnam. Thousands of Cambodians continued to enter Thailand in the 1980s, and by the end of the decade those in refugee camps were thought to exceed 300,000.

Sihanouk's government-in-exile

In 1982 an uneasy alliance was reached among the three groups opposing the Vietnamese-backed regime in Phnom Penh, and a government-in-exile was established with Sihanouk as president and Son Sann as prime minister. This government, despite United Nations recognition, received little support from Cambodians inside the country and was largely ineffectual. The member groups of the coalition continued independently to resist the Phnom Penh regime, with the larger and better-equipped forces of the Khmer Rouge being the most effective.

Cambodia since 1990. The political stalemate that developed among the four groups vying for power was broken in the late 1980s, when international political pressure, an economic boycott of Cambodia led by the United States, and a reduction in aid from the Soviet Union contributed to Vietnam's decision to withdraw its forces from Cambodia (completed in 1989). Freed from Vietnamese tutelage, the Phnom Penh government took two initiatives that increased its popularity: it legalized the ownership of real estate and officially encouraged the practice of Buddhism. The withdrawal of the Vietnamese also allowed the resistance factions to seek through negotiation the political objectives that they had been unable to obtain by military action against the Phnom Penh government; they were encouraged in this endeavour by their foreign patrons.

These negotiations, which had been conducted for some time and which had intensified after 1989, led in 1991 to two significant results. The first was the creation of a largely ceremonial coalition government—the Supreme National Council (SNC)—that contained representatives of all four factions. Although the SNC was recognized by the United Nations, effective control in most of Cambodia remained in the hands of the Phnom Penh regime. The second and more important result was the conclusion of a peace agreement among the factions that also provided for a popularly elected government. The UN Security Council, with the backing of the factions, endorsed this treaty and agreed to establish in the country a peacekeeping operation consisting of both soldiers and civil servants, which would monitor progress toward conducting elections, temporarily run several government ministries, and safeguard human rights. The operation was difficult to implement, notably because of the lack of cooperation by the Khmer Rouge. Free elections, however, were held in 1993, and Sihanouk returned as the country's monarch, heading a coalition government. In the mid-1990s the Khmer Rouge collapsed. In 1998 free elections again produced a coalition government, the country was admitted to ASEAN, and Pol Pot died. Subsequent UN efforts to bring surviving Khmer Rouge leaders before an international tribunal were hampered by the Cambodian government, which included some former Khmer Rouge members. (D.P.Ch.)

For later developments in the history of Cambodia, see the BRITANNICA BOOK OF THE YEAR.

Laos

Laos (in full, Lao People's Democratic Republic; Lao: Sathalanalat Paxathipatai Paxaxôn Lao) is a small, landlocked country constituting the northwestern portion of the Indochinese Peninsula. It is bounded on the north by China, on the northeast and east by Vietnam, on the south by Cambodia, on the west by Thailand, and on the northwest by Myanmar (Burma). Laos extends about 650 miles (1,050 kilometres) from northwest to southeast and has a total area of approximately 91,400 square miles (236,800 square kilometres). The capital is Vientiane (Lao: Viangchan).

PHYSICAL AND HUMAN GEOGRAPHY

Mountains

The land. *Relief.* Dominating the landscape of Laos are its inhospitable, forest-covered mountains, which in the north rise to a maximum elevation of 9,245 feet (2,818 metres) above sea level at Mount Bia and everywhere constitute an impediment to travel. The principal range lies along a northwest-southeast axis and forms part of the Annamese Cordillera (Chaîne Annamitique), but secondary ranges abound. Three notable landscape features of the interior of Laos may be mentioned. In the northern province of Xiangkhoang, the Plain of Jars (Lao: Thông Haihin; the name derived from large prehistoric stone jars discovered there) consists of extensive rolling grasslands rather than a true plain and provides a hub of communications. The central provinces of Bolikhamxay and Khammouan contain karst landscapes of caverns and severely eroded limestone pinnacles. Finally, in the south the Bolovens Plateau, at an elevation of about 3,600 feet, is covered by open woodland and has generally fertile soil. The only extensive lowlands lie along the eastern bank of the Mekong River.

Drainage. The general slope of the land in Laos is downhill from east to west, and all the major rivers—the Tha, Beng, Ou, Ngum, Kading, Bangfai, Banghiang, and Kong—are tributaries of the Mekong (Lao: Mènam Khong). The Mekong flows generally southeast and south along and through western Laos and forms its boundary with Myanmar and most of the border with Thailand. The course of the river itself is severely constricted by gorges in northern Laos, but, by the time it reaches Vientiane, its valley broadens and exposes wide areas to flooding when the river breaches its banks, as it did in August 1960. A few rivers in eastern Laos flow eastward through gaps in the Annamese Cordillera to reach the Gulf of Tonkin; the most important of these is the Ma River, which rises in Xiangkhoang province.

Bill Wassman/Apa Photo Agency SIN

The Ou River at Muang Khoua, northern Laos.

Soils. Soils in the floodplains are formed from alluvium deposited by rivers and are either sandy or sandy clay with light colours or sandy with gray or yellow colours; chemically, these are neutral to slightly acidic. Upland soils derived from crystalline, granitic, schistose, or sandstone parent rocks generally are more acidic and much less fertile. Southern Laos contains areas of laterite (leached and iron-bearing) soils, as well as basaltic soils on the Bolovens Plateau.

Climate. Laos has the typical tropical monsoon climate of the region, though the mountains provide some variations in temperature. During the rainy season (May to October), the winds of the southwest monsoon deposit an average rainfall of between 50 and 90 inches (1,300 and 2,300 millimetres), with totals reaching 160 inches on the Bolovens Plateau. The dry season (November to April) is dominated by the northeast monsoon. Minimum temperatures average between 60° and 70° F (16° and 21° C) in the cool months of December through February, increasing to highs of more than 90° F (32° C) in March and April, just before the start of the rains. In the wet season the average temperature is 80° F (27° C).

Plant and animal life. Laos has tropical rain forests of broad-leaved evergreens in the north and monsoon forests of mixed evergreens and deciduous trees in the south. In the monsoon forest areas the ground is covered with tall, coarse grass called *tranh;* the trees are mostly second growth, with an abundance of bamboo, scrub, and wild banana. The forests support a rich wildlife, including elephants, gaurs (wild oxen), deer, bears, tigers and leopards, monkeys, and a large variety of birds.

Settlement patterns. Laos is predominantly rural and agricultural. The numerous isolated valley communities preserve a variety of different traditions and dialects. Villages usually are located close to rivers and roads that give the people access to itinerant traders as well as to each other. Most villages are laid out around a main street or open area, farmlands being adjacent to the residential areas. Every village, if it can, has a Buddhist temple and supports at least one monk. The temple compound usually includes a public building that serves as a school and a meeting hall. Village leadership is usually divided, the headman having authority in secular matters and the monk in religious.

Typical village layout

The hill peoples usually are organized on tribal lines and live in smaller groupings. They are hunters and gatherers of forest products, as well as farmers, but their practice of shifting cultivation prevents them from establishing permanent villages. Hill peoples living close to the lowland areas tend to acquire the languages and cultures of their neighbours and to engage in limited trade with them; those living at higher elevations remain unacculturated.

Urban life in Laos is limited mainly to the capital, Vientiane, the former royal capital, Louangphrabang, and four or five other large towns. With the exception of Louangphrabang, all are located near the Mekong River in the floodplain area. Their populations are predominantly Lao, with smaller groups of Chinese, Vietnamese, and Indians. Compared with the cities of Thailand, Malaysia, or Vietnam, those of Laos are small and provincial.

The people. *Ethnic and linguistic characteristics.* The peoples of Laos are divided by language, culture, and location. Lao officials distinguish four basic ethnolinguistic groups: the Lao-Lum, or valley Lao; the Lao-Tai, or tribal Tai; the Lao-Theung, better known as the Mon-Khmer; and the Lao-Soung, or Hmong and Man. Mountain people sometimes are called Kha ("Slaves"), a pejorative term.

The Lao-Lum live in the lowlands, on the banks of the Mekong and its tributaries, and in the cities. They speak Laotian Tai, which is closer to the language spoken by the Thai of Thailand than it is to the language of the local Tai-speaking tribes.

The Lao-Tai include such local groups as the Black Tai (Tai Dam) and Red Tai (Tai Deng), both names referring to the dress of the women; the Tai Neua, or Tai of the north; the Tai Phuan of Xiangkhoang province; and the Phu Tai. The Lao-Tai live throughout the country, chiefly in upland areas, and their various dialects are mutually intelligible.

The Lao-Theung (Mon-Khmer) include many groups of people scattered throughout Laos, northeastern Myanmar, northern Thailand, and southern China. They are thought to be descendants of the earliest peoples to inhabit the region. These people do not form a single coherent group but rather include between 25 and 30 distinct groups, some of which are closely related while others are only tenuously identified as being part of this linguistic group.

The Lao-Soung, which include the Hmong (formerly called the Meo, or Miao) and the Man (Yao), are believed to have been coming from southern China since the late 18th century. They are divided into subgroups, and neither constitutes a large proportion of the population of Laos.

Other distinct linguistic groups are few in number. Speakers of Tibeto-Burman dialects, who also came from southern China, live in the north and northwest. Chinese and Vietnamese live primarily in the urban areas. Initially, French was the language of the Lao elite and of the cities, but by the 1970s English had begun to displace it. Under the leadership of the Lao People's Revolutionary Party, Vietnamese has become the third language of the elite.

Prior to the establishment of the Lao People's Democratic Republic (LPDR) in 1975, it was accurate to say that the Lao-Lum peoples had a distinct pattern of culture and dress. They also had a well-defined social structure, differentiating between royalty and commoners. The members of the elite included only a few outsiders who were not descendants of nobility. Most of the elite lived in the cities, drawing their incomes from rural land rents or from urban occupations. After 1975 a new elite emerged representing the victorious leftist forces. Many of this group, however, were of aristocratic origin.

Social and political structure

Traditionally, Lao-Tai society had a stratified social structure and a political hierarchy. The people were organized into groups larger than villages called *muong,* each of which was ruled by a hereditary ruler called the *chao muong.* Within this broad grouping, however, there were ethnic variations. Among the Black Tai, the nobility consisted of two descent groups, the Lo and the Cam, who provided the rulers of the *muong.* The religious leaders came from two other descent groups, the Luong and the Ka. The Black Tai tribal organization had three levels: the village; the commune, which was composed of a number of villages; and the overall *muong.* The latter two were ruled by nobles, while the village headman was selected from among the commoners by the heads of households. The Red Tai had a similar social structure, with the addition of a council of five to aid the *chao muong.* The nobility owned the land and had the right of service from the commoners.

The Mon-Khmer had no political or social structure beyond the village. They were led by a village headman, who was their link to the central government; but his role in the village was not clear.

Among the Lao-Soung, the Hmong maintained the tradition of a king and subchiefs and a large-scale organization, although in practice this usually was limited to the village. The village consisted of several extended families. In some villages, all the heads of households were members of a single clan, and the head of the clan was the headman of the village. Where several clans resided together in a large village there were several headmen, one being the nominal head and the link to the government. The headman had real authority in the village and was aided by a council. The Hmong extended their organization beyond the village for military purposes.

Religion. The predominant religion of Laos is Theravāda Buddhism, which is professed by most Lao and by a small number of other ethnic groups. Most of the rest of the people are animists, or spirit worshipers, especially in the more isolated upland areas. Many see no contradiction in being both, since Buddhism shows the way to enlightenment, while spirit worship helps a person to cope with daily and local problems. Among the hill peoples, especially those who have migrated from southern China, are found groups that mix Confucian ideas with Buddhism and animism. One subgroup of the Mon-Khmer, the Lamet, practices ancestor worship, and the Hmong are both spirit and ancestor worshipers. Roman Catholic and Protestant missionaries were present in the country before 1975, but only a tiny proportion of the population is Christian. The Vietnamese, who live both in the cities and in the northeastern rural areas, practice a mixture of Mahāyāna Buddhism and Confucianism.

Demographic trends. Laos is an underpopulated country. It has the lowest population density of any Southeast Asian nation, and its population is also one the most youthful. A high birthrate is offset by one of the highest infant-mortality rates in the region. About half the people are concentrated in the lowlands, and only about one-fifth are urban dwellers. The Lao-Lum are the largest ethnic group. There has been a considerable out-migration of people from Laos since the mid-1970s, including most of the country's educated and professional elite.

Migration

The economy. Laos is one of the world's poorest countries. The disruption during the civil-war period and the economic policies of the early years of the LPDR—notably the attempt to collectivize agriculture—resulted in economic stagnation in the country. By 1980, however, the government had begun to pursue more pragmatic development policies, and in 1986 it introduced market-oriented

reforms. Subsequently, private enterprise has been allowed to operate on every level, and foreign investment has been encouraged. A number of nongovernmental organizations, including some from the United States, have been assisting the government, mainly in the fields of rural development and public health.

Resources and power. Laos has a number of mineral resources, including coal, iron, copper, lead, gold, tin, gypsum, and precious stones. Tin has been mined commercially since colonial times, and gypsum has become important; the other minerals have been worked only in primitive and unsystematic ways.

Laos has considerable hydroelectric power potential. Electricity produced from a dam on the Ngum River north of Vientiane and sold to Thailand is one of the country's most valuable exports.

Agriculture, forestry, and fishing. The chief occupation of the people is agriculture, with the vast majority engaged in rice farming. In years with normal harvests, Laos is self-sufficient in rice production, a reflection of the success of private landownership and market incentives. In addition to a variety of food crops, modest amounts of such cash crops as sugarcane, tobacco, and coffee are produced. Agricultural production, however, is vulnerable to natural calamities, and Laos is subject to periodic droughts and floods, with both sometimes occurring in the same year. A significant proportion of rice production comes from upland dry rice raised by hill people using shifting-cultivation methods (*i.e.,* fields are cleared and cultivated for a few years before being abandoned and allowed to revert to forest). The government has considered this practice to be a major cause of deforestation and has attempted to resettle hill peoples on plains where they can adopt sedentary agricultural practices.

Another problem for the government has been the illegal cultivation of the opium poppy, mainly by Hmong hill people. Programs to curb production have had limited success because of the profitability of opium production and the inaccessibility of growing areas.

The extensive forests of Laos produce teak and other timber, benzoin (a balsamic resin), cardamom, and stick lac (used to make shellac). Much of the deforestation in the country has been blamed on logging operations and on the cutting of wood for fuel, which are believed to have caused the erosion of hillsides and the silting of rivers that has heightened the severity of droughts and floods. The regulation of logging by the government has been difficult because of poor communications and the high number of illegal operations.

Deforestation

Fishing is important for lowland dwellers, and aquaculture has increased. Livestock raising has grown in importance since 1980.

Industry. The main activities of the country's tiny industrial sector are food processing (rice milling and beverage production), sawmilling, and the manufacture of building materials and a variety of light consumer goods. Handicrafts also are important.

Finance and trade. Until the late 1980s, the government controlled all banking activities. Since then it has fostered the development of a private banking sector. Foreign investment and joint ventures with foreign companies have been officially encouraged.

The chief exports, in terms of value, are timber and other forest products, electricity, coffee, and tin. Major imports include foodstuffs, petroleum products, machinery, and transport equipment. The main trading partners are Thailand, Japan, China, and Hong Kong. Imports always have vastly exceeded exports in value, leaving a gap to be filled by foreign aid.

Transportation. A major obstacle to the economic and social development of Laos has been its poor transportation system. Rivers and roads are the major avenues of communication, supplemented by air transport. The Mekong River is the major north-south commercial artery; all but two sections of it—the Khone Falls and the rapids of Khemmarat (Khemarat)—are navigable for at least part of the year. Large barges ply the deeper sections of the rivers between towns, but most of the water traffic is carried in smaller craft. Some of the mountain people build bamboo rafts and float their goods to market, selling the raft along with the goods. During French rule, a primitive network of roads was created. The main artery joined Saigon (now Ho Chi Minh City, Vietnam) with Louangphrabang, and several lesser roads led eastward through mountain passes into Vietnam. During the Vietnam War, road building and improvement were undertaken by the United States, China, and what was then North Vietnam; the best-known of these works was the Ho Chi Minh Trail, a complex of roads, fords, and ferries in the Annamese Cordillera. Laos itself has no railway system, but Thailand's railways funnel goods and passengers to Laos. Small aircraft came into use during the war, and some inhabitants of Laos flew in airplanes before they had seen an automobile. Vientiane has an international airport.

Administration and social conditions. *Government.* Since its establishment in December 1975, the Lao People's Democratic Republic (LPDR) has been effectively controlled by the communist Lao People's Revolutionary Party. This party, in alliance with the Vietnamese communists, carried out the revolution that ended in its seizure of power and the abolition of the monarchy. Top government positions—beginning with the president, who is head of state, and the prime minister, who is also the party chairman—are held by high-ranking party members, who constitute a Central Committee with a Politburo at the head. A constitution adopted in 1991 provides for a National Assembly, the members of which are elected to five-year terms. The judicial system is headed by the Supreme People's Court.

The country is divided into 16 provinces—roughly from north to south, Phôngsali, Louang Namtha, Bokeo, Oudomxay, Louangphrabang, Houaphan, Xaignabouri, Vientiane, Xiangkhoang, Bolikhamxay, Khammouan, Savannakhét, Saravan, Xékong, Champasak, and Attapu—as well as Vientiane Municipality, the administrations of which are said to have considerable autonomy in economic matters. The provinces are subdivided into districts and villages. At each level of local government there are party committees and administrative committees, which often are headed by the same individuals.

Provinces

Security. Laos maintains a small military force consisting almost entirely of army personnel, with smaller air and naval branches. Internal security measures have been strictly enforced, as the regime fears political opposition linked to a large exile population and sporadic armed resistance within the country.

Education. Education has been reorganized since 1975. The government has set up a number of agricultural schools, sent teachers to provincial villages to provide literacy instruction, and opened new primary, secondary, and teacher-training schools. School enrollment and literacy rates subsequently have increased significantly from levels in the mid-1970s.

Health and welfare. Medical care in general is inadequate and unevenly distributed in Laos, with most of the health care facilities located in urban areas. Infant mortality rates are high, and life expectancy is low. Respiratory diseases, influenza, malaria, and gastroenteritis are the major health problems. The departure of most of the country's physicians after 1975 created a serious problem for the new government. It began to build village infirmaries and dispensaries in most of the provinces and to train medical workers. These village medical workers, often using only traditional medicinal herbs, now provide most of the country's primary health care.

Cultural life. *The arts.* The basis of Laotian culture is religion and tradition. Art, literature, music, and drama draw mainly from these sources. Towns along the Mekong are exposed to Western culture through Thai mass media.

Theravāda Buddhism entered the country in the 14th century. This religion and Hinduism have been major influences on cultural and intellectual life in Laos. The story of the Buddha and Hindu myths are the subjects of the carvings and sculptures found in all religious places. In the south, Khmer influences on the peoples of Laos are strong; in the north, Myanmar and Thai influences are readily apparent. As elsewhere in Southeast Asia, religious symbols, stories, and themes have been modified and lo-

calized. The snake, for example, representations of which adorn religious and royal buildings, symbolizes the benevolent spirit of the water and the protector of the king.

The Laotians have a variety of folk arts, including weaving, basketmaking, wood and ivory carving, and silverwork and goldwork. There are a number of Laotian musical instruments, of which the *khene*, a bamboo wind instrument, is most widely known. Music is not written down but is played from memory.

Dance

Dancing is a profession rather than a form of recreation; the professional dance troupes travel throughout the country performing for religious celebrations or on important holidays. Their main themes are drawn from the Indian epics. All professional dancers are male, the female roles being performed by young men and boys.

Laotian literature is predominantly religious and linked to the Buddhist tradition. There is also a secular literary stream based on themes of the Hindu epic poems, which have been transmuted into popular language; an example of this is the Laotian epic the *Sin Xay*, written between the mid-16th and the late 17th century. The popular poems and songs are often satiric.

Press and broadcasting. The government controls all aspects of the media. The largest-circulating daily newspaper is *Pasason*, published in Vientiane; it is the official organ of the ruling party. Also published in Vientiane is the quarterly journal *Aloun Mai*. The official news agency is Khaosan Pathet Lao (KPL). The National Radio of Laos broadcasts in a number of languages, principally Lao, English, and French. There is also a government-run television station. (J.Si./A.J.D.)

For statistical data on the land and people of Laos, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The Lao people, the predominant ethnic group in present-day Laos, are a branch of the Tai-speaking peoples who by the 8th century AD had established a powerful kingdom, Nanchao, in southwestern China. From Nanchao, the Tai gradually penetrated southward into the Indochinese Peninsula; their migration was accelerated in the 13th century by the Mongol invasions of southern China by Kublai Khan. The Lao, together with other Tai peoples, gradually supplanted various indigenous tribes (collectively known as Kha, or "Slaves") that from the 5th century on had lived in what is now Laos under the suzerainty of the Khmer empire of Cambodia. During the 12th and 13th centuries, the Tai established the principality of Muong Swa (later Luang Prabang, now Louangphrabang), which was ruled by various Tai leaders and the history of which survives in Laotian legend and myth.

The first Laotian state

Lan Xang. Recorded Laotian history begins with Fa Ngum, the ruler who founded the first Laotian state, Lan Xang ("Kingdom of the Million Elephants"), with the help of the Khmer sovereign at Angkor. Fa Ngum was a great warrior, and between 1353 and 1371 he conquered territories that included all of present-day Laos and much of what is today northern and eastern Thailand. He extended the Indo-Khmer civilization to the upper Mekong River and introduced Theravāda Buddhism, which had been preached by Khmer missionaries from Angkor.

In 1373 Fa Ngum was succeeded by his son Oun Hueun (reign name Sam Sen Thai), who did much to organize the pattern of administration and defense for the kingdom. After his death in 1416, a long period of calm—broken only by a Vietnamese invasion in 1479—allowed his successors to complete the work of organizing Lan Xang. This period of peace and tranquility ended with Photisarath (ruled 1520–48), who involved Lan Xang in a struggle against Myanmar and the Thai kingdom of Ayutthaya (Ayudhya) that lasted two centuries. Photisarath waged three wars against Ayutthaya and succeeded in placing his son Setthathirath on the throne of the Tai state of Chiang Mai (Chiengmai), marking Lan Xang's maximum territorial expansion. On Photisarath's death, Setthathirath returned to rule as Setthathirath I (ruled 1548–71). His reign was marked by the loss of Chiang Mai to the Myanmar, by the transfer of the capital from Luang Prabang to Vien Chan (now Vientiane), and by the repulsion of two Myanmar invasions that took place about 1565 and 1570.

When he died (1571), the Myanmar seized Vien Chan (1574) and ravaged the country, which lapsed into anarchy until Souligna Vongsa ascended the throne in 1637 and restored order. He fixed the frontiers with Vietnam and Siam (Thailand) by means of treaties and led two victorious expeditions against the principality of Chieng Khouang in the south. A defender of Buddhism and a patron of the arts, he embellished Vien Chan and made it a vibrant intellectual centre. His reign is considered by Laotians to be a golden age.

When Souligna Vongsa died in 1694, one of his nephews seized the throne with the help of a Vietnamese army, thus placing Lan Xang under Vietnamese rule and initiating a period of chaos that ended in the partition of the kingdom of Lan Xang. Other members of the royal family refused to accept Vietnamese vassalage. With the northern provinces under their control, they declared themselves independent (1707) and established the separate kingdoms of Luang Prabang and Vien Chan. The south seceded in turn and set itself up as the kingdom of Champassak (1713). Split into three rival kingdoms, Lan Xang ceased to exist.

Under foreign rule. During the 18th century, the three Laotian states, which were continually at loggerheads, tried to maintain their independence from the Myanmar and the Siamese, both of whom were contending for the control of western Indochina. The weakness of these states, resulting from their disunity, inevitably caused them to fall prey to the Siamese.

Siamese control

Vien Chan, which had sided with the Myanmar, was invaded (1778), annexed, and made a state subject to Siam (1782). Luang Prabang, which had supported the Siamese, was invaded by the Myanmar (1752), who imposed their rule upon it until the Siamese supplanted them (1778). In the south, Champassak, which had supported a Myanmar revolt against the Siamese, also was invaded (1778) and transformed into a dependency of Siam. Each of these kingdoms was placed under the control of a Siamese commissioner. The kings of Champassak, Vien Chan, and Luang Prabang were allowed to rule in their respective kingdoms but had to pay tribute to Bangkok. Their appointments to the throne were made in Bangkok.

The last king of Vien Chan, Chao Anou (also called Anouvong; ruled 1805–28), attempted to shake off this yoke. First, he strengthened the bonds of allegiance uniting Vien Chan to Vietnam (1806), whose influence in Indochina had grown to rival that of Siam. Next, he persuaded Bangkok to give his son the governorship of Champassak, thus extending his frontiers as far as the old southern boundaries of Lan Xang. Thinking that the British, who had just defeated the Myanmar, were going to attack Siam, he led three armies against Bangkok (1827). The Siamese, however, regrouped their forces, marched on Vien Chan, and defeated Anou, who fled to Vietnam. Vien Chan was pillaged and destroyed. In 1828 Anou attempted another attack but was again defeated. Vien Chan was made a Siamese province.

For the Siamese, the annexation of Vien Chan was the first step toward the creation of a great empire. They next extended their colonization of the eastern bank of the Mekong to protect themselves from an eventual Vietnamese expansion westward. They therefore garrisoned Champassak (1846) and Luang Prabang (1885) and stationed troops as far as the Annamese Cordillera. Siamese expansion toward the northeast—where the mountain states were placed under the cosuzerainty of Vietnam and Luang Prabang—provoked the protests of the French, who had established a protectorate over Vietnam. France entered into negotiations with Bangkok (1886) to define the Siamese-Vietnamese frontier and won the right to install a vice-consul in Luang Prabang. The office was entrusted to Auguste Pavie, who, owing partly to his popularity with the Laotians, succeeded in winning Luang Prabang over to France. After a number of Franco-Siamese incidents in the Mekong River valley, French ships made a show of strength off Bangkok in 1893. Later that year, on the advice of the British, Siam withdrew from the eastern bank of the Mekong and gave official recognition to the French

protectorate in the evacuated territory. French annexation was completed by treaties with Siam (called Thailand from 1939) in 1904 and 1907.

French protectorate

The French organized this territory as a protectorate, with its administrative centre at Vientiane, and allowed it autonomy in local matters. The kingdom of Luang Prabang survived, but the other provinces were placed under the direct authority of a French official. France paid little attention to Laos until the Japanese invaded the Indochinese Peninsula during World War II; in 1941, under Japanese pressure, the Vichy government restored to Thailand the territories acquired in 1904. In March 1945 the Japanese took outright administrative control of Indochina from the French, and the following month the independence of Laos was proclaimed.

Two movements sprang up at this time. The first was anti-Japanese and was represented by the court of Luang Prabang and Prince Boun Oum of Champassak; the second was anti-French (the Free Laos movement, or Lao Issara), was located in Vientiane, and was led by Prince Phetsarath. These two movements remained in conflict until the return of French troops, which in early 1946 compelled the supporters of the Lao Issara to flee to Thailand. France, in a temporary agreement, recognized the internal autonomy of Laos under the king of Luang Prabang, Sisavang Vong. Finally, after a constitution was promulgated and general elections were held, a Franco-Laotian convention was signed in July 1949 by which Laos was granted limited self-government within the French Union. All important power, however, remained in French hands.

Although many of the Lao Issara leaders were prepared to work with the French under this new arrangement, their decision was opposed by a more radical group led by Kaysone Phomvihan and Prince Souphanouvong. Under Souphanouvong's presidency, a new political movement, the Pathet Lao ("Land of the Lao"), was proclaimed (1950) that joined forces with the Viet Minh of Vietnam in opposing the French. The Pathet Lao remained unreconciled when the French took further steps toward granting independence to Laos in October 1953, while still retaining control of all military matters in the kingdom. Between 1950 and early 1954, the Pathet Lao gained strength in northeastern Laos and had a firm grip on two of the country's provinces when the peace conference in Geneva brought the First Indochina War to an end.

(P.-B.L./M.E.O.)

Laos after the Geneva Conference, 1954–75. The Geneva Accords of 1954 marked the end of French rule on the Indochinese Peninsula. The participating nations (including France, Great Britain, the United States, China, and the Soviet Union) at the Geneva Conference agreed that all of Laos should come under the rule of the royal government and should not undergo partition (as did Vietnam). The agreements, however, did provide for two "regroupment zones" in provinces adjacent to what was then North Vietnam to allow the Pathet Lao forces to assemble. This resulted in the de facto control of these areas by the Laotian communists, while the rest of the country was ruled by the royal government.

The uneasy peace in Laos was short-lived, as hostilities broke out between leftist and rightist factions in 1959. Another conference in Geneva was convened in May 1961, culminating in an agreement in July 1962 that called for the neutralization of the country and the formation of a tripartite government. The new government consisted of factions from the left (the Pathet Lao, who were linked to North Vietnam), the right (linked to Thailand and the United States), and neutrals (led by Prince Souvanna Phouma). Once again, however, the cease-fire was brief. The coalition split apart by 1964, and the larger war centred in Vietnam engulfed Laos. In this expanded war, Laos, like Cambodia, was viewed by the major protagonists as a sideshow.

Tripartite government

The agreement negotiated by the United States and North Vietnam at Paris in 1973 called for a cease-fire in each of the countries of the Indochinese Peninsula, but only in Laos did peace actually occur. In February 1973, just a month following the agreement, the Laotian factions signed the Vientiane Agreement, which provided again for a cease-fire and yet another coalition government composed of factions from the left and right, presided over by Souvanna Phouma. As political control in Vietnam subsequently tipped toward the communists, following the American departure there, the Pathet Lao gained political ascendancy in Laos. When the communists marched into Saigon (now Ho Chi Minh City, Vietnam) and Phnom Penh, Cambodia, the right-wing forces in Laos lost heart and most of their leaders fled, permitting a bloodless takeover by the Laotian communists in mid-1975. The Laotian communists proclaimed an end to the 600-year-old monarchy in December 1975 and established the Lao People's Democratic Republic (LPDR).

The Lao People's Democratic Republic. Guiding the politics of the newly established republic was the Lao People's Revolutionary Party (LPRP), the communist party of Laos. Its politburo was dominated by a small, cohesive band of revolutionaries who had founded the party in 1955 (called the Lao People's Party until 1972) and had engaged in persistent revolutionary activity until their takeover in 1975. These leaders had a long and intimate relationship with their Vietnamese communist allies. Prior to founding the party, they had been members of the Vietnamese-led Indochina Communist Party. Most spoke Vietnamese, and some had family ties with Vietnam. The party's general secretary (until 1992), Kaysone Phomvihan, had a Vietnamese father; second-ranked Nouhak Phoumsavan and third-ranked Prince Souphanouvong had Vietnamese wives. Their worldview had been shaped by their shared revolutionary struggle with Vietnam. Moreover, the Vietnamese had numerous channels—party, military, and economic—through which they directly conveyed their influence. Thus, the new state was intimately linked to Vietnam and closely followed its policy line until the late 1980s.

In the early years of the LPDR, the leadership declared its twin economic goals to be "socialist transformation with socialist construction." Following the Vietnamese communist model, the party leaders attempted to create agricultural collectives in the countryside and to nationalize the limited industry and commerce in the towns. Former members of the Royal Lao Army as well as of the deposed government—perhaps as many as 30,000—were incarcerated in "reeducation" camps. These and other repressive political measures and the grim economic conditions in Laos compelled some 10 percent of the country's population to flee across the Mekong River to Thailand after 1975.

Reforms in the 1980s

Beginning in the early 1980s and continuing with greater conviction after 1986, party leaders introduced significant changes in both domestic and foreign policy. Within Laos, market incentives were adopted, and government economic enterprise was decentralized. Private investment and joint ventures also were encouraged. To the relief of Lao peasants, attempts at collectivization were abandoned in favour of family-operated farms. Although the ruling party retained unchallenged control, as in Vietnam, both political freedom and participation were enlarged. A new constitution was promulgated in 1991. Citizens were permitted to move about their country more freely and even to cross the Mekong River to Thailand with fewer impediments.

The most significant change in foreign affairs was the reduction of Vietnamese political control and military presence. Concurrently, Lao relations with Thailand improved, and there was a warming of relations with China. Border-demarcation agreements were signed with China, Myanmar, and Vietnam. Relations with the United States, which had become strained but were never severed, also improved. After the establishment of the LPDR, the Soviet Union and the countries of eastern Europe became the country's primary trading partners and donors of economic assistance. The dissolution of the Soviet Union and transformation of its European allies, however, has obliged Laos to turn increasingly for aid to Japan, Australia, Sweden, the European Community, and international organizations.

(J.J.Z.)

For later developments in the history of Laos, see the BRITANNICA BOOK OF THE YEAR.

Malaysia

The Southeast Asian nation of Malaysia consists of two dissimilar regions: Peninsular, or West, Malaysia on the Malay Peninsula and East Malaysia on the island of Borneo. The capital is Kuala Lumpur, located on the western side of Peninsular Malaysia. Malaysia has a total area of 127,584 square miles (330,442 square kilometres), which includes about 265 square miles of inland water. Of this total, Peninsular Malaysia constitutes about 50,810 square miles and East Malaysia about 76,510 square miles.

Peninsular Malaysia occupies most of the Malay Peninsula south of latitude 6°40′ N. To the north it is bordered by Thailand, with which it shares a land boundary of some 300 miles (480 kilometres). To the south, at the tip of the peninsula, is the island republic of Singapore, with which Malaysia is connected by a causeway. To the southwest, across the Strait of Malacca, is the Indonesian island of Sumatra.

East Malaysia consists of the states of Sarawak and Sabah and is separated from Peninsular Malaysia by some 400 miles of the South China Sea. These two states occupy most of the northwestern coastal part of the large island of Borneo and share a land boundary with the Indonesian portion (Kalimantan) of the island. Within Sarawak is a small coastal enclave containing the sultanate of Brunei.

Malaysia, a member of the Commonwealth, represents the political marriage of territories that were formerly under British rule. When it was established on Sept. 16, 1963, Malaysia was composed of Malaya (now Peninsular Malaysia), Singapore, Sarawak, and Sabah. In August 1965 Singapore seceded from the federation and became an independent republic.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The long, narrow, and rugged Malay Peninsula extends to the south and southwest from Myanmar and Thailand. The Malaysian portion of it is about 500 miles long and—at its broadest east-west axis—about 200 miles wide. About half of Peninsular Malaysia is covered by granite and other igneous rocks, one-third is covered by stratified rocks older than the granite, and the remainder is covered by alluvium. At least half the land area is more than 500 feet (150 metres) above sea level.

Peninsular Malaysia is dominated by its mountainous core, which consists of a number of roughly parallel mountain ranges aligned north-south. The most prominent of these is the Main Range, which is about 300 miles long and has peaks rising to elevations of more than 7,000 feet. Karst landscapes—limestone hills with characteristically steep, whitish gray sides, stunted vegetation, caves created by the dissolving action of water, and subterranean passages—are distinctive landmarks in central and northern Peninsular Malaysia. Bordering the mountainous core are the coastal lowlands, 10 to 50 miles wide along the west coast of the peninsula but narrower and discontinuous along the east coast. Settlement and development have taken place primarily along the west coast.

East Malaysia is an elongated strip of land approximately 700 miles long with a maximum width of about 170 miles. The coastline of 1,400 miles is paralleled inland by a 900-mile land boundary with Kalimantan. For most of its length, the relief consists of three topographic features. The first is the flat coastal plain. In Sarawak, where the coastline is regular, the plain averages 20 to 40 miles in width, while in Sabah, where the coastline is rugged and deeply indented, it is only 10 to 20 miles wide. Inland from the coastal plain is the second topographic feature, the hill-and-valley region. Elevations there generally are less than 1,000 feet, but isolated groups of hills reach heights of 2,500 feet or more. The terrain in this region is usually irregular, with steep-sided hills and narrow valleys. The third topographic feature is the mountainous backbone that forms the divide between East Malaysia and Kalimantan. This region, which is higher and nearer the coast in Sabah than in Sarawak, is composed of an eroded and ill-defined complex of plateaus, ravines, gorges, and mountain ranges. The summits of the ranges are between 4,000 and 7,000 feet. Mount Kinabalu, at 13,455 feet (4,101 metres) the highest peak in Malaysia, towers above this mountain complex.

Features of East Malaysia

Drainage. Peninsular Malaysia is drained by an intricate system of rivers and streams. The longest river—the Pahang—is only 270 miles long. Streams flow year-round because of the constant rains, but the volume of water transported fluctuates with the localized and torrential nature of the rainfall. In the western part of the peninsula such heavy rainfalls may occur at any time of year, but in the eastern part they are more likely to occur during the northeast monsoon (November to March). Prolonged rains often cause floods, especially in areas where the natural regimes of the rivers have been disrupted by uncontrolled mining or agricultural activities.

As in Peninsular Malaysia, the drainage pattern of East Malaysia is set by the interior highlands, which also form the watershed between Malaysia and Indonesia. The rivers, also perennial because of the year-round rainfall, form a dense network covering the entire region. The longest river in Sarawak, the Rajang, is about 350 miles long and is navigable by shallow-draft boats for about 150 miles from its mouth; its counterpart in Sabah, the Kinabatangan, is of comparable length but is navigable for only about 120 miles from its mouth. The rivers are important because they provide a means of communication between the coast and the interior. Settlement also has taken place along the rivers, as it did on the peninsula in an earlier period.

Soils. The soils of both portions of Malaysia have been exposed for a long period of time to intense tropical weathering, with the result that most of their plant nutrients have been leached out. Soils typically are strongly acidic and coarse-textured and have low amounts of organic matter. Any organic matter when exposed to weathering is rapidly oxidized, and the soils consequently become even poorer. Soil erosion is always a danger on sloping ground, where such additional measures as building contour embankments or planting protective cover crops are required.

Only a small proportion of the soils of Peninsular Malaysia are fertile; regular applications of fertilizers are therefore necessary in order to sustain crop yields. Generally, soil conditions in Sarawak and Sabah do not differ greatly from those on the peninsula. Of these three regions, only Sabah has appreciable areas of fertile soils. These are found in particular in the southeastern coastal areas, where the parent material from which the soil is formed is composed of chemically basic volcanic materials.

Climate. Both peninsular and insular Malaysia are in the same latitudes and are influenced by similar airstreams. They consequently have high temperatures and humidities, heavy rainfall, and a climatic year patterned around the northeast and southwest monsoons. The country is influenced by eight or nine major airstreams flowing from the northeast, the south, and the west; the advance and retreat of these airstreams are responsible for the division of the climatic year into four seasons. These are the northeast monsoon (from November or December until March), the first intermonsoonal period (March to April or May), the southwest monsoon (June to September or early October), and the second intermonsoonal period (October to November). The onset and retreat of the two monsoons are not sharply defined.

Seasons

Malaysia has an equatorial climate, but the narrowness and topographic configuration of each portion—central mountainous cores with flat, flanking coastal plains—facilitate the inland penetration of maritime climatic influences. In addition, the monsoons further modify the climate. The northeast monsoon brings heavy rain and rough seas to the exposed coasts of southwestern Sarawak and northern and northeastern Sabah. The southwest monsoon, however, affects mainly the southwestern coastal belt of Sabah. Floods are common, especially along the west coast of Sabah. Neither peninsular nor insular Malaysia is in the typhoon belt, but their coasts occasionally are subject to the heavy rainstorms associated with squalls.

Temperatures are uniformly high throughout the year. On the peninsula, they average 78° to 82° F (25° to 28° C) for most lowland areas. In coastal areas in East Malaysia, minimum temperatures range from 72° to 76° F (22° to 24° C), and maximum temperatures from 88°

to 92° F (31° to 33° C); temperatures are lower in the interior highland regions. The mean annual rainfall on the peninsula is approximately 100 inches (2,540 millimetres); the driest location, Kuala Kelawang (formerly Jelebu), near Kuala Lumpur, receives about 65 inches of rain per year, while the wettest, Maxwell's Hill, northwest of Ipoh, receives some 200 inches annually. Mean annual rainfall in Sabah varies from 80 to 140 inches, while most parts of Sarawak receive 120 inches or more per year.

Plant life. The characteristic vegetation of Malaysia is dense, evergreen rain forest. Rain forest still covers about half of the peninsula and some three-fourths of Sarawak and Sabah; another fraction is under swamp forest. Soil type, location, and altitude produce distinctive vegetation zones: tidal swamp forest on the coast, freshwater- and peat-swamp forest on the ill-drained parts of the coastal plains, lowland rain forest on the well-drained parts of the coastal plains and foothills up to an altitude of about 2,000 feet, and submontane and montane (lower mountain-type) forest above that elevation. The highly leached and sandy soils of parts of central Sarawak and the coast support an open, heathlike forest known locally as *kerangas* forest.

The flora of the Malaysian rain forest is among the richest in the world. There are some 8,000 species of flowering plants, of which at least 2,500 are trees. An acre (0.4 hectare) of forest may have as many as 100 different species of trees, as well as shrubs, herbs, lianas (creepers), and epiphytes (nonparasitic plants that grow on other plants and derive nourishment from the atmosphere). The forest canopy is so dense that little sunlight can penetrate it. As a result, the undergrowth usually is poorly developed and—contrary to popular belief—is not impenetrable. Much of the original rain forest has been destroyed by severe wind and lightning storms, by indigenous peoples clearing it for shifting cultivation, or by clearances made for agricultural or commercial purposes. When such cleared land is subsequently abandoned, coarse grassland, scrub, and secondary forest develop.

Loss of the original forest

Animal life. The forests and scrublands are inhabited by a large variety of animal life. Mammals on the peninsula include the elephant, tiger, seladang (or Malayan gaur, a massive wild ox), Sumatran rhinoceros, tapir (a hoofed and snouted quadruped), wild pig, and many species of deer, including the *pelandok,* or chevrotain (a small, deer-like ruminant). Crocodiles, monitor lizards, and cobras also are indigenous to the country, while the green sea turtle and the giant leathery turtle nest regularly on the beaches of the east coast.

Animal life in East Malaysia is even more varied than it is on the peninsula. In addition to the peninsular species, East Malaysia is also the home of the fast-disappearing orangutan and rhinoceros, the sun bear (also called the honey bear), and the unique proboscis monkey—a reddish tree-living species. There also are vast numbers of cave swifts, whose nests are regularly collected and sold as the main ingredient of bird's nest soup.

Settlement patterns. The people of Malaysia are predominantly rural. Their settlements are similar in appearance and pattern to those of their rural counterparts elsewhere in Southeast Asia. The basic unit in both East and Peninsular Malaysia is the kampong (village, or community of houses), consisting of dwellings on stilts.

The houses of Peninsular Malaysia usually are built of wood, and traditionally they have a thatched roofing called atap that is woven from the leaves of the nipa palm (a species also used for basketry); increasingly common are roofs of corrugated metal. Each house is surrounded by a grove of coconut palms and banana, papaya, and other fruit trees. The four main types of Malay settlement—fishing villages, paddy (wet-rice) villages, cash-crop villages, and mixed-crop villages—despite their variations, conform to the same basic pattern. Most other rural settlements on the peninsula are associated with peoples who have settled in the country since the early 19th century. The earliest of these were the mining camps, which sprang up in the tin fields in the west. Some have since grown into large towns, but others—especially in the Kinta River valley—still remain small. The British introduced the plantation system of agriculture, and the subsequent cultivation of rubber and the oil palm changed the face of rural Peninsular Malaysia. Added to the landscape was the plantation, or estate, settlement, typically a group of buildings consisting of the processing factory and storehouse, the labourers' quarters, and the manager's house.

New Villages represent a type of settlement that is unique to Peninsular Malaysia. These originally were simply groups of buildings that were established as defensive sites near roads between 1948 and 1960, during the Emergency, the formal name for the period when the British administration was engaged in suppressing the communist guerrilla uprising. With the end of the Emergency in 1960, some of the New Villages were abandoned, but most of them became permanent settlements. A more recent and significant government program has involved the resettlement of poor Malays into forest areas, which are cleared and planted in rubber trees and oil palms; since the mid-1950s, more than 100,000 families have been resettled.

Rural resettlement

About three-fourths of the population of East Malaysia is still rural, and it is in the rural areas that the greatest variety of settlement types is encountered. This variety is a direct reflection of the considerable ethnic diversity of the population and of the fact that indigenous as well as immigrant groups are settled in the rural areas. The non-Malay indigenous ethnic groups—including the Iban (Sea Dayak), Bidayuh (Land Dayak), Kenyah, Kayan, and Murut—are thinly scattered in the foothill country and, to some extent, in the coastal lowlands as well. They are primarily shifting cultivators and live in locations on or near riverbanks. Their traditional dwelling is the longhouse, which is more commonly found in Sarawak than in Sabah. Each longhouse is raised on stilts and is composed of a number of rooms, known as *bileks*; each *bilek* houses a family. A longhouse can grow by accretions of related families, and an Iban longhouse may in time reach a length of 40 or more *bileks*. Some groups, such as the Melanau of Sarawak and the Kadazan of Sabah, have abandoned the longhouse settlement form, adopting instead the single-family dwelling of the Malays.

The Malays and Melanau of East Malaysia share many common characteristics with their rural counterparts on the peninsula. They tend to be riverine and coastal peoples, with an economy based on agriculture and fishing. Many live in kampongs set in the midst of coconut palms, mangroves, or other swamp trees. Their houses generally are built on stilts. The Melanau live in the large delta swamp region between Bintulu and Rajang. The rural Chinese in Sarawak have settled in the region between the coast and the uplands, usually in homesteads strung along both sides of the roads, where they grow cash crops in smallholdings. Their houses are commonly built at ground level and thus are easily distinguishable from the stilt-raised dwellings of the indigenous peoples.

The cities and large towns of Peninsular Malaysia were built up during the colonial and postcolonial periods and are distributed mainly in the tin and rubber belt along the west side of the peninsula. The towns are associated with mining, purchasing, processing, distributing, exporting, and administrative functions, and each town usually performs several of these functions. Some towns are located at coastal or riverine sites, emphasizing the early importance of water transport, while more modern towns have been built in inland areas served by road, rail, and air transport.

There is a growing number of satellite towns such as Petaling Jaya (outside Kuala Lumpur), although most of the towns of Peninsular Malaysia are unplanned, having grown up around small nuclei. Urban land use generally is mixed, and buildings are put to multiple uses. Streets, built for a more leisurely era, are narrow and often congested. In the larger centres, such as Kuala Lumpur, Ipoh, and George Town (Pinang, or Penang), distinct central business districts similar to those in Western cities have emerged. These are characterized by heavy population and traffic densities, high land values, and a concentration of shopping, banking, insurance, entertainment, and other facilities.

Urbanization in East Malaysia has proceeded slowly. Only a small percentage of people live in towns. The

largest towns are Kuching, Sibu, and Miri in Sarawak and Sandakan, Kota Kinabalu, and Tawau in Sabah. The large towns invariably are located on coastal or riverine sites. The layout and appearance of these towns are markedly similar: a wharf area, rows of Chinese shop-houses in the central business districts, more substantial buildings in the governmental administrative area, and one or more timber and atap kampongs built on the riverbanks.

The people. The population of Malaysia is unevenly divided between Peninsular and East Malaysia, with the vast majority living in Peninsular Malaysia. The population shows great ethnic, linguistic, cultural, and religious diversity. A significant distinction is made between indigenous peoples (aborigines and Malays, collectively often called *bumiputra*) and immigrants (primarily Chinese and South Asians). In addition, there are important differences among the indigenous peoples themselves and among religious groups.

Peninsular Malaysia

Ethnic composition, languages, and religions. The Malay Peninsula, situated at one of the great maritime crossroads of the world, has long been the meeting place of peoples from other parts of Asia. As a result, the population shows the ethnographic complexity typical of Southeast Asia as a whole. In general, there are four groups of people, given in the order of their appearance on the peninsula: the Orang Asli (aborigines), the Malays, the Chinese, and the South Asians. In addition, there are small numbers of Europeans, Americans, Eurasians, Arabs, and Thai.

The Orang Asli constitute the smallest group and can be divided ethnically into the Jakun, who speak an archaic Malay, and the Semang and Senoi, who speak languages of the Mon-Khmer language family. They are primarily adherents of traditional religions, but a number have been converted to Islām.

The Malays originated in different parts of the peninsula and archipelagic Southeast Asia. They constitute about two-thirds of the population and are politically the most important group. They share with each other a common culture, speak a common Austronesian language—Malay (officially called Bahasa Malaysia), which is the national language—and are overwhelmingly Muslim. Adherence to Islām is regarded as one of the most important factors distinguishing a Malay from a non-Malay, and the number of Malays who are not Muslim is negligible. Minor differences in dialect, culture, and physical characteristics are noticeable among the Malays living in the south in Johor state, on the east coast in the states of Kelantan and Terengganu, and on the west coast in the states of Negeri Sembilan, Perak, Kedah, and Perlis.

The Chinese, who make up about one-third of the peninsular population, originally migrated from southeastern China. They are ethnically homogeneous but are less homogeneous than the Malays in language and religion. Several different dialects are spoken, notably Hokkien, Cantonese, Hakka, and Hainanese. Thus, it may be necessary for two Chinese to converse in Mandarin Chinese, English, or Malay. A minority, the Baba Chinese, speak a Malay patois, although otherwise they remain Chinese in customs, manners, and habit. The Chinese do not have a dominant religion; most of them, while subscribing to Confucian moral precepts, are either Buddhist or Taoist. A small minority is Christian.

The peoples from South Asia—Indians, Pakistanis, and Tamils from Sri Lanka—constitute about 10 percent of the population of Peninsular Malaysia. Linguistically, they can be subdivided into speakers of Dravidian languages (Tamil, Telugu, Malayālam, and others) and speakers of Indo-European languages (Punjābī, Bengali, Pashto, and Sinhalese). Numerically, the Tamil speakers are the largest group. Most of the Indians and Sri Lankans are Hindu, while the Pakistanis are predominantly Muslim. Some Indians have been converted to Christianity. The Sikhs, from the Punjab, adhere to their own religion, Sikhism.

East Malaysia

The population of East Malaysia is ethnographically even more complex than that of Peninsular Malaysia. The government has tended to oversimplify the situation in Sarawak and Sabah, officially recognizing only some of the dozens of ethnolinguistic groups in those two states. The main ethnic groups in Sarawak are the Chinese, various speakers of mutually unintelligible Austronesian languages including the Iban (Sea Dayak), the Malays, the Bidayuh (Land Dayak), and the Melanau.

The Chinese of Sarawak, like those on the peninsula, originally came from southeastern China. The relative size of each dialect group is reversed, however, as speakers of Hakka and Fu-chou (Hokchiu) in Sarawak outnumber those speaking Cantonese and Hokkien. As in Peninsular Malaysia, nearly all the Chinese of Sarawak follow Confucianism and practice Buddhism or Taoism.

The Iban are the largest and most important indigenous group in Sarawak. Their origins are obscure, but traditionally they were headhunters. The Iban are a homogeneous people speaking a language described as a type of pre-Islāmic Sumatran Malay. Most of them live in the interior uplands, where they are longhouse dwellers practicing shifting cultivation. They have a distinctive culture, in which nearly every activity is influenced or governed by their animist religious beliefs.

The Malays of Sarawak are a heterogeneous group of people, among whom only a few are of peninsular origin. Most are the descendants of aboriginal peoples who since the mid-15th century have converted to Islām and adopted the Malay way of life. Although ethnically diverse, they are culturally homogeneous, speaking a common language and practicing Islām.

The Bidayuh live in hill country, most being found in the far western portion of Sarawak. Although all are of the same ethnic group, they speak a number of different but related dialects that to some extent are mutually intelligible. The majority of the Bidayuh practice traditional religions, but Christian missionaries have made some converts among them.

The Melanau differ ethnically from the Sarawak Malays, but their dialects, which are distinct from Malay, do not differ sufficiently to constitute a barrier to communication. The great majority of Melanau are Muslim, with the rest (except for a small number of Christians), following traditional religions. Other indigenous peoples—including the Kenyah, Kayan, Kedayan, Murut, Kelabit, Bisaya (Bisayah), and Punan—contribute much to Sarawak's ethnic and cultural diversity.

Sabah also has a kaleidoscopic mixture of peoples. The largest groups are the Kadazan, Chinese, Bajau, and Murut, while a significant proportion consists of such indigenous peoples as the Kedayan, Orang Sungei, Bisaya, Sulu, and Tidong. Europeans, Eurasians, Malays, Indonesians, Filipinos, and South Asians make up the remainder.

Kadazan society consists of a number of tribes, each speaking a dialect that the others can understand. The great majority of Kadazan are animists, although a significant proportion are Christian and a small number are Muslim. Most of the Chinese are Hakka-speaking, the other important dialects being Cantonese, Hokkien, Teochew, and Hainanese. The Bajau are not a cohesive community, as they are split into two main groups: sedentary agriculturists living on the north coast and those who live by the sea on the east coast. Most are Muslim, but not all of them can communicate with each other. The Murut of Sabah are descended from the same people as the Kadazan and are ethnically different from the Murut of Sarawak. They are shifting cultivators. Although they are divided into subtribes, their languages are mutually intelligible. Most follow traditional religions, with a significant minority being Christian.

Demographic trends. The average life expectancy in Malaysia has increased significantly since the end of World War II. Death rates for all groups are less than half of what they were in the late 1950s, and infant mortality rates also have declined sharply. Mortality rates tend to be lower in towns and cities, where there are better health services; since the Chinese usually are urban dwellers, their death rates are lower than those for the country as a whole. Birth rates have declined significantly since the 1960s, but the rate of natural increase has remained high.

Before World War II, there was a free flow of people to and from both Peninsular and East Malaysia, and the rate of population growth was greatly influenced by a net

surplus from in-migration. A series of laws passed since 1945, and particularly after the political separation from Singapore, now restricts the entry of immigrants from all countries. Thus, immigration is no longer a major cause of population growth.

Population density

The major area of population concentration in Peninsular Malaysia is an axis of economic development on the west side of the peninsula. Much smaller concentrations are found in the Kelantan and Terengganu river deltas in the northeast. The remainder of the peninsula—the interior uplands and most of the east—generally is sparsely populated. Slightly more than half the population of the peninsula's urban centres is Chinese, and about one-third is Malay; Indians and Pakistanis make up most of the remainder.

The population density of East Malaysia is considerably less than that for the peninsula. As in the west, the main concentrations are along the coasts and rivers. In Sarawak the heavy concentration of people in the southwest makes this region the most important in East Malaysia. The population is similarly clustered on the coast in Sabah, but riverine settlements are less important than in Sarawak. As on the peninsula, the urban population is predominantly Chinese.

The economy. Malaysia's economy has been transformed since 1970 from one based primarily on the export of raw materials (rubber and tin) to one that is among the strongest, most diversified, and fastest-growing in Southeast Asia. Primary production remains important: the country is the world's largest producer of rubber and palm oil, exports considerable quantities of petroleum and natural gas, and is one of the world's largest sources of commercial hardwoods. Increasingly, however, Malaysia has emphasized export-oriented manufacturing to fuel its economic growth. Using the comparative advantage of a relatively inexpensive but educated labour force, well-developed infrastructure, political stability, and an undervalued currency, Malaysia has attracted considerable foreign investment, especially from Japan and Taiwan.

The focal point of this growth has been the manufacture of electrical and electronic products and textiles, which together have become one of the most important sources of export earnings. The success of the manufacturing effort has been reflected by the development of a variety of heavy industries, including steelmaking and automobile assembly—the latter implemented through a Malaysian-Japanese joint venture. Peninsular Malaysia, especially the urban area of Kuala Lumpur and the rest of the developed area along the western side of the peninsula, accounts for nearly all of the country's manufacturing output.

The New Economic Policy

Since the early 1970s the Malaysian government has championed a social and economic restructuring strategy, first known as the New Economic Policy (NEP), that seeks to strike a balance between the goals of economic growth and the redistribution of wealth. Traditionally, the Malaysian economy has been dominated by the country's Chinese and South Asian minorities. The goal of the NEP has been to endow the Malays and other indigenous groups with greater economic opportunities and to develop their management and entrepreneurial skills. Official economic policy also has encouraged the private sector to take a greater role in the restructuring process. A major component of this policy has been the privatization of many public-sector activities, including the national railway, airline, automobile manufacturer, and telecommunications company.

Malaysia's systems of public finance—auditing and organization of accounts, parliamentary control, and revenue collection—are generally based on British principles. The primary role of the country's fiscal system is to raise revenue for governmental expenditure, rather than being a mechanism to manipulate the pace of economic activity, the level of employment, or prices. The greater part of government revenues are raised by taxation—roughly equally divided between direct (income) taxes and indirect taxes (*e.g.*, customs and excise duties).

Malaysia's rapid economic expansion has created a great demand for additional labour for the manufacturing and service sectors. The labour shortage has tended to increase wages. Nonetheless, there has been a relatively limited flow of workers from East to Peninsular Malaysia despite the economic incentives, prompting interest in recruiting foreign workers.

Resources. Malaysia is rich in mineral resources. The major metallic ores are tin, bauxite (aluminum), copper, and iron. A host of minor ores found within the country include manganese, antimony, mercury, and gold. The production of tin formed one of the main economic pillars upon which the country's development effort has been built. It is found largely in alluvial deposits along the western slopes of the Main Range in Peninsular Malaysia, with smaller deposits on the east coast of the peninsula. Malaysia's most valuable mineral resources, however, are its reserves of petroleum and natural gas. The major fields are all offshore, off the east coast of the peninsula and off Sarawak. Malaysia also has large reserves of coal, peat, and wood, and it has considerable hydroelectric potential.

Agriculture. Agriculture, forestry, and fishing were the traditional basis of the Malaysian economy. Their contribution to the nation's gross domestic product (GDP) gradually has declined from roughly one-third in 1970 to less than one-fifth. These three activities still engage the largest percentage of the workforce, but the proportion is diminishing.

The main food crop, rice, is grown on small farms. Despite the widespread advances brought about by the introduction of improved plant varieties and chemical fertilizers and pesticides (the so-called Green Revolution), rice production has declined. The main causes of this have been unfavourable weather conditions and the loss of farm labour to urban manufacturing jobs, the latter situation having the effect of reducing the amount of land that can be worked. As a result, the country is not self-sufficient in rice production and must make up the shortfall with imports, chiefly from Thailand. Shifting cultivation is practiced primarily in East Malaysia.

Cash crops

The most important cash crops are palm oil and rubber. Together these account for a significant (though declining) proportion of Malaysia's commodity exports. The production of palm oil and rubber has been subject to considerable fluctuations in the price of these commodities, which has resulted in a decline in the number of plantations. Palm oil has become more important than rubber in terms of value. Also important are cocoa, pepper, and coconuts.

Forestry and fishing. The extensive forests of both Peninsular and East Malaysia are heavily exploited for their timber. The lowland evergreen tropical rain forest is the principal forest formation of commercial importance, being rich in species of the economically valuable Dipterocarpaceae family. Sarawak and Sabah account for the greater part of all timber production. Concern has been raised, however, about the pace of deforestation caused by the combination of shifting agriculture and intensive logging operations in East Malaysia. Attempts have been made to curtail log exports from the region and to substitute wood-based industries, such as the manufacture of plywood and furniture. Logging remains important in Peninsular Malaysia, although much of the easily accessible timber has been cut. The region also has a long history of careful forest management and conservation, and the effects of deforestation there have not been as serious.

Traditionally, most of Malaysia's fish catch has been from the shallow seas off its coasts, where the water's nutrient levels—and hence its productivity—generally have been low. In the 1970s the country's fishing industry was modernized, notably by the addition of trawlers and mechanized fishing boats. This allowed the more abundant offshore fish resources to be tapped, leading to a dramatic increase in catches. Malaysia has become a major fishing nation, even though production peaked in about 1980 and much of the fishing industry has remained confined to the overexploited shallow onshore waters. Aquaculture production also has increased, although the country's potential has remained largely undeveloped.

Mining and power. Extractive industries still contribute significantly to Malaysia's GDP. Tin output has declined dramatically since the 1970s, however, because of the depletion of readily accessible alluvial deposits, rising mining

costs, and fluctuating demand in the world tin market. Petroleum production has increased in importance, as has the production of liquefied natural gas; together they account for a major portion of the country's commodity export earnings. Of some importance are copper, bauxite, and iron, although production of these ores has varied greatly with fluctuations in world markets. Iron output has declined as high-grade deposits have been depleted. Malaysia's bauxite production is centred near Johor at the south end of the peninsula, while the nation's copper comes from western Sabah.

Malaysia's petroleum resources constitute the major energy source for power generation. The country's proven reserves of coal and peat are not economical to mine and have remained largely unexploited. Wood and charcoal were the traditional domestic fuels, but in the urban areas they have been displaced by bottled gas. Hydroelectricity accounts for more than one-fourth of all energy production, but most of this generating power is concentrated on the peninsula; the abundant rainfall and steep gradients of the rivers in the interior highlands offer good opportunities for further exploitation in both portions of the country.

Manufacturing

Industry. Manufacturing has undergone rapid expansion since the 1970s and has become the leading edge of Malaysia's economic growth. It now accounts for the largest share of the country's GDP, although primary activities still employ more workers. Growth has been especially notable in industries assembling electronic equipment, electrical machinery, and appliances and those making chemical products and textiles. The main development goal has been the manufacture of goods for export, with a lesser emphasis on import substitution. One strategy designed to promote manufactured exports has been the establishment of a number of free trade zones, which have provided duty-free access to imported raw materials and semifinished parts and numerous investment and export incentives. Industrial estates also have been established in less-developed parts of the country to stimulate manufacturing and to balance industrial growth, but manufacturing capacity has remained highly concentrated. The country's heavy industries—more important politically than economically—generally have been saddled with excess capacity and high production costs. Increasingly, development strategy has shifted to the promotion of small and medium industries that manufacture their own parts and acquire technology from more economically developed countries, the aim being to move beyond the stage of assembly-only manufacturing.

Finance and trade. Malaysia has an active and growing financial sector, which has been encouraged by government policies that promote foreign investment, market competition, and the privatization of publicly held enterprises. Banking and insurance are regulated by the state-run Bank Negara Malaysia. The state permits a variety of banking activities, including a semipublic bank that operates on Islāmic financial principles. Kuala Lumpur has a commodities exchange and a stock exchange.

Malaysia's export structure has shifted dramatically since 1970, from one dominated by rubber and tin to one in which manufactured goods now account for more than half of all export earnings. Electrical and electronic products constitute the largest proportion of exported manufactures. Commodities exports, however, remain important. Imports are dominated by machinery and other manufactured goods. Malaysia's chief trading partners are Japan, Singapore, and the United States. Such newly industrialized Asian countries as South Korea and Taiwan account for a growing share of trade. Malaysia is a member of the Association of Southeast Asian Nations (ASEAN), and trade with other ASEAN nations (outside of that with member Singapore) also is increasing.

Transportation. Malaysia's transportation systems have been improved considerably since independence, although demand generally has outstripped capacity. In addition, much more attention has been given to developing the infrastructure of Peninsular Malaysia than that of East Malaysia. The peninsula's road network includes high-speed express highways and numerous hard-surfaced secondary roads; it is especially well-developed in the major

Roads

An electronics factory in the free trade zone on Penang Island, Malaysia.

Milt and Joan Mann/Cameramann International

industrial states. The road network in East Malaysia is much less extensive, with fewer paved roads. Malaysia's small railway system, confined primarily to the peninsula, is of much less significance than its roads.

River transport is of great importance in East Malaysia, especially in Sarawak. Malaysia's long and accessible coastlines have long fostered maritime trade. Several ports, notably George Town and Port Kelang on the Strait of Malacca, have become major container-handling facilities. Numerous other ports have been developed, including Kuantan on the eastern coast of the peninsula, Kuching in Sarawak, and Kota Kinabalu in Sabah. Air transport has grown rapidly, with passenger traffic increasing especially on the peninsula. An internal air network connects all Malaysian states, and Kuala Lumpur and Pinang have international airports.

Administration and social conditions. *Government.* Malaysia is a federal constitutional monarchy with a nonpolitical head of state, or *yang di-pertuan agong* ("paramount ruler"), who is elected from among nine state hereditary rulers for a five-year term. The federal legislature consists of the Senate (Dewan Negara) and the House of Representatives (Dewan Rakyat). The federal government also has a prime minister and cabinet, an independent judiciary, and a politically neutral civil service.

The powers of the federal parliament are relatively broad and include the authority to legislate in matters concerning government finances, defense, foreign policy, internal security, the administration of justice, and citizenship. The state legislatures, however, retain responsibility for issues pertaining to Islāmic law and for matters regarding personal and family laws affecting Muslims, as well as for land laws. The constitution also provides that some issues may be addressed either by the federal or by a state legislature.

The House of Representatives functions in a manner similar to that of the British House of Commons. It has a membership of 180, of which 132 are from Peninsular Malaysia, 27 from Sarawak, and 21 from Sabah. Members are elected to office from single-member constituencies by a simple majority to terms of five years. The Senate has a membership of 69; of these, 43 members are appointed by the *yang di-pertuan agong* on the recommendation of the prime minister (including 2 from the federal territory of Kuala Lumpur and 1 from the federal territory of Labuan), and the other 26 are elected—2 from each of the 13 states—by the state legislative assemblies. Voting in either house is by a simple majority, but amendments to the constitution require a two-thirds majority. A bill passed by both houses and sanctioned by the *yang di-pertuan agong* becomes a federal law.

The *yang di-pertuan agong* appoints a prime minister

from the members of the House of Representatives. On the advice of the prime minister, the *yang di-pertuan agong* then appoints the other ministers who make up the cabinet. The number of ministers is not fixed, but all must be members of the federal parliament.

Local rule

Each state of Malaysia has its own written constitution, legislative assembly, and executive council responsible to the legislative assembly and headed by a chief minister. Several Malay states—Johor, Kedah, Kelantan, Pahang, Perak, Selangor, and Terengganu—have hereditary rulers (sultans). The raja (king) is the ruler in Perlis, and the *yang di-pertuan besar* ("chief ruler") in Negeri Sembilan. The heads of state of Melaka, Pulau Pinang (Penang Island), Sarawak, and Sabah—known as *yang di-pertuan negeri*—are appointed to office. The ruler of a state acts on the advice of the state government. The constitution provides for parliamentary elections and for elections to state legislatures, to be held at least every five years.

Political parties. Malaysia has a multiparty political system, and since about 1970 it has held free elections and changed prime ministers peacefully. Party affiliation generally is based on ethnicity, though less so than at independence. Malaysian political life is dominated by the National Front (Barisan Nasional), a broad coalition of ethnically oriented parties that long has been controlled by the United Malays National Organization. The main opposition parties are the Democratic Action Party (consisting primarily of Chinese), the Muslim Unity Movement (a coalition of pro-Islāmic parties), and the Sabah People's Union. The Communist Party of Malaya and the offshoot Malaysian Communist Party are illegal opposition parties.

Justice. The constitution of Malaysia, which is the supreme law of the country, provides that the judicial power of the federation shall be vested in two High Courts, one in Peninsular Malaysia and the other in East Malaysia, and also in subordinate courts. Above the High Courts is the Supreme Court (Mahkamah Agung), with jurisdiction to hear and determine appeals from decisions by any High Court. The supreme head of the judiciary is the lord president of the Supreme Court.

Each High Court consists of a chief justice and a number of other judges—up to 33 in Peninsular Malaysia and up to 8 in East Malaysia. The High Court has unlimited criminal and civil jurisdiction and may pass any sentence allowed by law. Below the High Court are the subordinate courts, which consist of the Sessions Courts and the Magistrates' Courts. Both these lower courts have criminal and civil jurisdiction—criminal cases coming before one or the other court depending on the seriousness of the offense and civil cases depending on the sum involved. In addition, there are religious courts in those Malay states that are established under Islāmic law. These courts are governed by state—not federal—legislation.

Armed forces. The Malaysian armed forces have increased in strength and capability since the formation of Malaysia in 1963. After the withdrawal of British military forces from Malaysia and Singapore at the end of 1971, a five-nation agreement between Malaysia, Singapore, New Zealand, Australia, and Great Britain was concluded to ensure defense against external aggression. The ASEAN also provides additional regional security.

The armed forces consist of an army, navy, and air force. The army is the most experienced and the largest of the three, constituting about four-fifths of all military personnel. The Royal Malaysian Navy concentrates mainly on defending the long indented coastlines and narrow waters of the country. The Royal Malaysian Air Force has combat aircraft, as well as many transport aircraft and helicopters.

The states of Malaysia inherited from their common colonial past an internal security system based on the British model. The police force is well trained and combats not only crime but also armed insurrections.

Education. The federal government provides free, noncompulsory primary and secondary education. Primary-school enrollment is nearly universal on the peninsula but is lower in Sabah and Sarawak. The number of students advancing to the secondary level has increased considerably. Institutions of higher learning include the University of Malaya in Kuala Lumpur, the University of Science, Malaysia, in Pinang, the National University of Malaysia in Bangi, and the International Islāmic University in Petaling Jaya. Enrollment in higher education also has increased, and many Malaysian students have studied abroad.

Universities

Health and welfare. The general level of health has improved considerably since World War II, which has contributed significantly to the decline in death and infant-mortality rates. The country is free from many of the diseases that plague tropical countries, although such diseases as malaria are still a problem in rural areas. Health conditions and health facilities vary among the component states, being better in Peninsular Malaysia than in Sabah and Sarawak. Health services generally are better in the towns and cities than in the rural areas. Segments of the rural population continue to rely on traditional rather than modern medicine for treatment. Most of the modern health services are provided by the government. Welfare services, however, are provided by both government and private agencies and include relief programs for poor, elderly, and handicapped individuals.

The multicultural character of the population of Malaysia is visibly reflected in the wide variety of houses, which range from the traditional longhouses and stilt houses of the rural peoples to examples of modern high-rise architecture in the cities. Housing shortages are rare in rural areas, but squatter settlements are common in the larger towns and cities. A governmental housing authority has had success in establishing low-cost housing in urban areas.

Certain groups of people, especially in Sarawak and Sabah, live by hunting, gathering, fishing, and simple farming, thereby reducing somewhat the number of wage earners in the total economically active population. Because of the increasing pressure of population on the land, however, there is a growing tendency for young people to seek employment in manufacturing. Since wages in the manufacturing sector are significantly higher than those in agriculture, labour shortages continue to prevail in the rural economy. Industrialization has drawn increasing numbers of workers from the countryside to the cities and has created a greater demand for skilled workers.

Cultural life. Malaysia is a melting pot of several major cultural traditions that stem from archipelagic Southeast Asia as well as from China, South Asia, the Middle East, and the West. Malay culture and Bornean culture are indigenous to the area. In the first one and a half millennia AD, indigenous Malay culture in the Malay Peninsula and in other parts of Southeast Asia was strongly marked by pre-Islāmic Indian and early Islāmic influences. Indian contact with the Malay Peninsula extended from about the 2nd or 3rd to the late 14th century, exerting a profound influence on religion (Hinduism and Buddhism), art, and literature. Islām, introduced to Malacca (now Melaka) in the 15th century, soon became the dominant religion of the Malays. The introduction of Western cultural influences in the 19th century affected many aspects of Malay life, especially in technology, law, social organization, and economics. Contemporary Malay culture is thus multifaceted, consisting of many strands—animistic, early Hindu, early and modern Islāmic, and, especially in the cities, Western—and the collective pattern is distinct from other cultures and recognizably Malay.

Outside influences

Unlike the early Chinese traders who settled in Malacca and George Town (now Pinang) and were partially assimilated (at least to the extent of adopting the Malay language), the Chinese who emigrated in large numbers to the Malay Peninsula in the late 19th and early 20th centuries were usually transients who established self-contained communities. Chinese cultural influence has consequently been minimal. The Chinese immigrants themselves, moreover, did not form a homogeneous group. Their culture in Malaysia has its roots in the culture and civilization of prerevolutionary China, with modifications brought about by local circumstances and environment.

Most of the Indians and Pakistanis originally came as labourers to work in the coffee and rubber plantations. Like the Chinese, they also were mainly transients (until World War II), living in closed communities and remaining virtually unassimilated.

The communities of Malaysia have been affected by British colonial rule and Western cultural influences, especially in education and institutional forms. Traditions and cultural institutions have been least affected in the rural areas—in eastern Peninsular Malaysia and in the interior of East Malaysia—while the cities have been the focus of the most rapid cultural changes.

The arts. External cultural influences have made the least impact in music, dancing, literature, and the decorative arts. In East Malaysia the indigenous cultural background includes no written history or literature. Architecture is little developed, and the principal art forms are dancing and handicrafts, represented notably by the textiles handwoven by the Punan tribe, cloth made by the Bajau people, patterned rattan mats and basketwork, and wood carvings. Particularly on the peninsula, the artistic manifestations of Malay culture are mainly in literature, music, dancing, and the decorative arts. Painting and sculpture are poorly developed, primarily because Islām does not encourage the representation of the human form. Examples of Malay decorative arts include batik cloth (cloth hand-dyed by using a special technique), silverware, the handmade kris (a short sword or heavy dagger with a wavy blade), wood carving, and basketwork. Malaysian Chinese culture is derived from Chinese civilization and is represented by literature, drama, music, painting, and architecture. Some Malaysian artists—of Malay, Chinese, and Indian origin—also have begun to produce new, synthesized, and distinctively Malaysian art forms, especially in painting and architecture.

Press and broadcasting. The newspapers are all privately owned (many by political parties) and vary greatly in circulation, quality of reporting, and news coverage. Among the educated groups, the press is the principal source of information. The radio is relied on in remote rural areas. Television, however, is the most popular medium among all language groups. (O.J.B./T.R.L.)

Radio and television

For statistical data on the land and people of Malaysia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Because the Malay Peninsula links mainland and archipelagic Southeast Asia and because Malaysia is characterized by a geographic division, the history of the present-day country can be understood only within a regional framework and as part of the wider context of the western archipelago zone. The Strait of Malacca (Melaka) bisects this realm and long has constituted a crossroads for peoples, cultures, and trade passing through or taking root in the area. Influences from China, India, the Middle East, and, later, Europe followed the maritime trade. Malaya (Peninsular Malaysia) and Sarawak and Sabah (East Malaysia) have shared many historical patterns, but each region also has developed in unique ways.

The rise of Indianized states. Malaysia's prehistory remains insufficiently studied, but bone and artifact discoveries at the Niah Caves site in northern Sarawak confirm modern human habitation in the region that may date to 40,000 years ago. The vast cave complex contains an almost unbroken succession of human frequentations and occupations, including a unique sequence of the evolution of stone tools that persisted until some 1,300 years ago. Malaya has been inhabited for at least 6,000 to 8,000 years, archaeologists having unearthed evidence of Stone Age and early Bronze Age civilizations; Neolithic culture was apparently well-established by 2500 to 1500 BC. Traditional historiography postulated that successive waves of peoples, who now are the modern Malays, migrated into the region from China and Tibet during the 1st millennium BC, pushing earlier inhabitants into the western Pacific or remote mountain enclaves. More recently it has been suggested that instead the southward migration consisted of small groups who imposed their culture and language and created new ethnic fusions.

Small Malayan kingdoms appeared in the 2nd or 3rd century AD, a time when Indian traders and priests began traveling the maritime routes and brought with them Indian concepts of religion, government, and the arts. Over many centuries the peoples of the region, especially the royal courts, synthesized Indian and indigenous ideas, making brilliant and selective use of Indian models—including Hinduism and Mahāyāna Buddhism—in shaping their political and cultural patterns. The most significant complex of Indianized temple ruins has been found around Kedah Peak in northwestern Malaya. The peninsula and northern Borneo both lacked broad, fertile plains and were unable to support the pattern of densely populated classical Southeast Asian civilizations that flourished in Cambodia and Java. Although knowledge is scant and is based chiefly on Chinese written sources, it does appear that perhaps 30 small Indianized states rose and fell in Malaya, mostly along the east coast, during the 1st millennium AD. The most important of these states, Langkasuka, controlled much of northern Malaya. Malaya developed an international reputation as a source of gold and tin, populated by renowned seafarers. Between the 7th and 13th centuries many of these small, often prosperous peninsular maritime trading states may have come under the loose control of Śrīvijaya, the great Sumatra-based empire. At various times the Cambodian Angkor and Javanese Majapahit empires and the Tai Ayutthaya (Ayudhia) kingdom also claimed suzerainty in the region. The early states left a living legacy, traces of which can still be found in the political ideas, social structures, rituals, language, arts, and cultural practices of Malay Muslims.

Langkasuka

Although development was slower in more remote, less fertile northern Borneo, Sarawak had entered the Iron Age by AD 600. Archaeological excavations in the Sarawak River delta reveal many ancient sites containing evidence of both ironworking and an enormous trade with China and the Southeast Asian mainland. The local peoples exchanged edible bird's nests, rhinoceros horns, hornbill ivory, camphor, spices, wood, and other goods for Chinese ceramics, metal, and probably clothing. Neolithic boatbuilders along the east coast of Sabah already were involved in extensive interregional trade at about the same time; the maritime peoples of the area called the territory the "land below the wind" because it lay south of the typhoon belt.

The advent of Islām. By the late classical period a new religion, Islām, was filtering quietly into the region from outside, carried chiefly by Arab and Indian merchants. From the 13th through 17th century Sunnite Islām spread widely, coming from the Middle East via India. It offered an egalitarian message that challenged the power of the traditional elites and a complex theology that held much appeal for peasants and merchants in the coastal regions. The spread of Islām was intimately linked to the florescence of the great Indian Ocean maritime trading routes that connected China through the Strait of Malacca to India, the Middle East, and East Africa. Over these routes Indonesian spices, Malayan gold, and Chinese silks and tea traveled to Europe, sparking interest there in reaching the sources of these riches.

The arrival of Islām coincided with the rise of the great port of Malacca (now Melaka), established on Malaya's southwest coast by Sumatran exiles about 1400. The Indianized king—who sensibly and successfully sought a tributary relationship with powerful China—converted to Islām, becoming a "sultan" and hence attracting Muslim merchants. Soon Malacca became Southeast Asia's major trading entrepôt, while at the same time it gained suzerainty over much of coastal Malaya and eastern Sumatra. Malacca served as the main centre for the propagation of Islām as well as the eastern terminus of the Indian Ocean trading network. At its height in the late 15th century, Malacca hosted some 15,000 merchants from many countries, including Chinese, Arabs, Persians, and Indians; there were said to be more ships in the harbour than in any other port in the known world, attracted by a stable government and a policy of free trade. The Chinese admiral Cheng Ho called at the port several times in the first decades of the 15th century as part of the great Ming naval expeditions to the western Indian Ocean. Malacca's political and religious influence reached its height under Tun Perak, who served as prime minister (1456–98) after defeating the expanding Thai in a fierce naval

Malacca

Malacca empire in 1500.

battle; during his service Islām became well-entrenched in such districts (and subsidiary sultanates) as Johor, Kedah, Perak, Pahang, and Terengganu.

The mostly Islāmicized people of the Malacca area began calling themselves "Malays" (a likely elite reference to earlier Śrīvijayan origins). Thereafter, the term Malay applied to those who practiced Islām and spoke a version of the Malay language; identity and behaviour, rather than descent, became the criteria for being Malay, so that previously animist and Hindu-Buddhist peoples of various origins could identify themselves (and even merge) with the prestigious Malays. Over time a loose cultural designation became a coherent ethnic group spread throughout Malaya, northern and western Borneo, eastern Sumatra, and the smaller islands in between, a region that can be termed the "Malay world." Islām, however, came to overlay the earlier beliefs, so that, before the rise of religious reform movements in the 19th century, few Malays were orthodox Muslims. Hindu-influenced ritual remained important for the elite, and animist spirits were richly incorporated into Islāmic folk beliefs.

Early European intrusions. The fame of Malacca as the crossroads of Asian commerce had reached Europe by the beginning of the 16th century. The Portuguese, who for a century had been seeking a sea route to the Orient, finally arrived at Malacca in 1509, inaugurating a new era of European activity in Southeast Asian history. Although many Southeast Asians, including the inhabitants of northern Borneo, experienced little Western impact before the 19th century, Malaya was one of the first regions disrupted. In 1511 a Portuguese fleet led by Afonso de Albuquerque captured Malacca. Since fewer merchants chose to endure the high taxes and the conquerors' intolerance of Islām, Malacca languished under Portuguese control. Indeed, the undermanned Portuguese barely repulsed repeated assaults by the dynamic sultanate of Aceh (Acheh) of northern Sumatra. Aceh had leaped into the political vacuum created by Malacca's downfall; during the 16th and early 17th centuries Aceh was deeply involved in peninsular affairs, warring against various sultanates and at times controlling some or most of them. The Dutch, who replaced the Portuguese as the dominant European power in Southeast Asia, seized Malacca in 1641; they tried to revive trade, but the city never recovered its earlier glory.

Aceh

An emphasis on the often complex politics of the peninsula in the post-Malacca years obscures other significant developments. Sultanates continued to be created throughout the Malay world. Usually they were situated at the mouth of a major river and sought to control trade to and from the interior, which often was populated by seminomadic peoples such as the aboriginal Orang Asli of Malaya and the various indigenous peoples of Borneo. Newer, dynamic sultanates—such as Riau-Johor, Kedah, and Brunei (on Borneo's northwest coast)—took over some of the trading functions of Malacca and flourished for several centuries. Islām reached Sarawak and Sabah in the 15th and 16th centuries; many coastal peoples converted, but the interior remained largely animist until the 20th century. Malay political control spread, with the Brunei sultans laying claim to much of what is today Sarawak and Sabah—although their actual power seldom reached much beyond the coastal zone. Attempts by Brunei to control the interior often failed, especially after the aggressive, head-hunting Iban people commenced their migrations into Sarawak from western Borneo (16th through 18th centuries). The Siamese came to control some of the northern Malay sultanates, and the southernmost part of present-day Thailand still has a predominantly Malay Muslim population. The Malay sultanates included many, often feuding chiefdoms, and wars between—or within—sultanates occasionally erupted. Europeans considered the sultanate system politically unstable, but it reflected a chronic situation in which states constituted hierarchical but fluctuating spheres of influence that ruled over mobile populations.

During the 17th century many Minangkabau people migrated from western Sumatra into southwestern Malaya, bringing with them a matrilineal sociocultural system by which property and authority descended through the female side. They elected their chiefs from among eligible aristocratic candidates, a model that has been incorporated into contemporary Malaysia's selection of a king. Later the Minangkabau formed a confederation of nine small states (Negeri Sembilan). The political pluralism of Malaya in the 18th century also facilitated large-scale penetration of the peninsula by the Buginese, a people from southwestern Celebes (Sulawesi) with a well-earned reputation as maritime traders. Buginese immigrants established the sultanate of Selangor in the mid-1700s and also gained dominance in the vigorous sultanate of Johor, at the southern tip of the peninsula, a prosperous trading entrepôt that attracted Asian and European merchants. Despite continuous movement of peoples from the archipelago into the area, Malaya and northern Borneo remained sparsely populated into the early 19th century. Many present-day Malays are descendants of immigrants from elsewhere in archipelagic Southeast Asia who arrived after 1800. Indeed, immigrants from Java, Celebes, and Sumatra demonstrated a tendency to merge with the existing Malay community over time, a process that steadily accelerated with the rise of Malay nationalism and vernacular education in the 1930s. Some of the customs brought by Minangkabau, Javanese, and other immigrants are still practiced in districts where they settled, contributing to the many regional variations of Malay culture and language.

Origins of the modern Malays

The colonization of Malaysia. *Malaya.* Except for Malacca, there was little Western influence in Malaya and northern Borneo until the late 18th century, when Britain became interested in the area. The British sought a source for goods to be sold in China, and in 1786 the English East India Company acquired Penang (or Pinang) Island, off Malaya's northwest coast, from the sultan of Kedah. The island soon became a major trading entrepôt with a chiefly Chinese population. British representative Sir Stamford Raffles occupied Singapore Island off the southern tip of the peninsula in 1819, acquiring trading rights in 1824; a strategic location at the southern end of the Strait of Malacca and a fine harbour made Singapore the centre for Britain's economic and political thrust in the peninsula. The British attracted Chinese immigrants to the sparsely populated island, and soon the mainly Chinese port became the region's dominant city and a major base for Chinese economic activity in Southeast Asia. By then the major industrial capitalist power in Europe, Britain next obtained Malacca from the Dutch in 1824 and thereafter governed the three major ports of the Strait of Malacca, which collectively were named the Straits Settlements. The British Colonial Office took direct control in 1867.

With the opening of the Suez Canal in 1869, the full effect of European technological superiority swept over Southeast Asia. The feuding Malay states were little prepared, with the exception of Johor, which was led by the

modernizing sultan Abu Bakar. The other state administrations generally were weak and failed to cope with their mounting problems, including the steady immigration of Chinese. By the early 19th century, the Chinese—who were being driven to emigrate by increasing poverty and instability in their homeland—began settling in large numbers in the sultanates along the peninsula's west coast, where they cooperated with local Malay rulers to mine tin. The Chinese organized themselves into tightly knit communities and formed alliances with competing Malay chiefs, and Chinese factions fought wars with each other for control of minerals. Chinese settlers also established towns like Kuala Lumpur and Ipoh, which later grew into major cities. The Chinese and Malays increasingly became leading elements in an inadequately integrated sociopolitical structure, a framework that produced chronic communal friction.

British intervention

British investors were soon attracted to Malaya's potential mineral wealth, but they were concerned about the political unrest. As a result, local British officials began intervening in various Malayan sultanates by the 1870s, establishing political influence (sometimes employing force or the threat of force) through a system of British residents (advisers). Initial intervention into Malayan internal affairs was crude and incompetent; the first British resident to Perak was murdered by Malays outraged at his assertive actions. Gradually, the British refined their techniques and appointed more able representatives; notable among these was Sir Frank Swettenham, who in 1896 became the first resident-general of a Malay federation of Perak, Selangor, Negeri Sembilan, and Pahang, with Kuala Lumpur as the capital. By 1909 the British had pressured Siam into transferring sovereignty over the northern Malay states of Kedah, Terengganu, Kelantan, and Perlis. Johor was compelled to accept a British resident in 1914. These sultanates remained outside the federation. Britain had now achieved formal or informal colonial control over nine sultanates, but it pledged not to interfere in matters of religion, customs, and the symbolic political role of the sultans. The various states kept their separate identities but were increasingly integrated to form British Malaya.

Sarawak. Sarawak's history also entered a new stage when the English adventurer James (later Sir James) Brooke intervened in a revolt against Brunei control and was appointed raja (governor) of the Sarawak River basin in 1841 by the Brunei sultan. Brooke inaugurated a century of rule by a remarkable English family and a new form of imperial endeavour. Simultaneously traditional Bornean potentates, benevolent autocrats, and cautious modernizers, the Brookes viewed themselves as protectors of Sarawak's people. Brooke spent his final years consolidating his control of surrounding districts and defending his government against various challenges. Although the first raja's political and financial position was often precarious, Sarawak eventually acquired the status of an independent state under British protection. Relations with Britain, however, were often strained, chiefly because of a consistent Brooke policy of incorporating territory at the expense of the declining Brunei sultanate. The present boundaries of Sarawak were achieved by 1906, but by then the once-powerful Brunei also had become a British protectorate.

Sabah. Sabah (North Borneo) was the last region to be brought under British control. In the early 1700s Brunei transferred its claims over much of Sabah to the sultan of Sulu, but, except in the northeast, actual Sulu power remained limited. Occasional resistance to Brunei or Sulu influence, as well as extensive coastal raiding and confusion of suzerainty, invited Western interest beginning in the 18th century. Despite short-lived American activity in the 1860s, British power proved most decisive. Britain had already acquired the offshore island of Labuan from Brunei by 1846. They gained a toehold in Sabah proper in 1872 when a British merchant, William Cowie, founded an east-coast settlement at Sandakan on lease from Sulu.

British North Borneo Company

By 1881 the British had obtained rights to much of Sabah and launched the British North Borneo Company, which, based in Sandakan, ruled the British protectorate from 1881 to 1941. The company operated the state in the interest of its shareholders but was only moderately prosperous, because of high overhead and poor management; its 60 years of rule, however, established the economic, administrative, and political framework of modern Sabah.

The impact of colonialism. The British presence in the region reflected several patterns: direct colonial rule in the Straits Settlements, more indirect control in some of the east-coast Malay sultanates, and family or corporate control in Borneo. Regardless of the political form, however, British rule brought profound changes, transforming the various states socially and economically. The Brookes and the North Borneo Company faced prolonged resistance before they consolidated their control, while occasional local revolts punctuated British rule in Malaya as well. In Sarawak in 1857, for example, interior Chinese gold-mining communities nearly succeeded in toppling the intrusive James Brooke before being crushed, while Muslim chief Mat Salleh fought expanding British power in Sabah from 1895 to 1900. The Brookes mounted bloody military campaigns to suppress head-hunting and to forcibly incorporate the autonomy-loving Iban, and similar "pacification" campaigns were carried out in Sabah. Those who resisted British annexation or policies were portrayed by colonialists as treacherous, reactionary rebels; but many of them are now hailed in Malaysia as nationalist heroes.

British administration eventually achieved peace and security. In Malaya the Malay sultans retained their symbolic status at the apex of an aristocratic social system, although they lost some of their political authority and independence. British officials believed that the Malay peasants needed to be protected from economic and cultural change and that traditional class divisions should be maintained. Hence, most economic development was left to Chinese and Indian immigrants, as long as it served long-term colonial interests. The Malay elite enjoyed a place in the new colonial order as civil servants. Many Malayan and Bornean villagers were affected by colonial taxes, however, and were forced to shift from subsistence to cash-crop farming; their economic well-being became subject to fluctuations in world commodity prices. Much economic growth occurred; British policies promoted the planting of pepper, gambier (a plant producing a resin used for tanning and dyeing), tobacco, oil palm, and especially rubber, which along with tin became the region's major exports. Malaya and British North Borneo developed classic extractive, plantation-based economies oriented to the resource and market needs of the industrializing West.

British authorities in Malaya devoted much effort to constructing a transportation infrastructure, in which railways and road networks linked the tin fields to the coast; port facilities also were improved to facilitate resource exports. These developments stimulated growth in the tin and rubber industries to meet world demand. The tin industry remained chiefly in immigrant Chinese hands through the 19th century, but more highly capitalized, technologically sophisticated British firms took over much of the tin production and export by World War II. The rubber tree was first introduced from Brazil in the 1870s, but rubber did not supersede the earlier coffee and gambier plantings until near the end of the century. By the early 20th century thousands of acres of forest had been cleared for rubber growing, much of it on plantations but some farmed by smallholders. Malaya became the world's greatest exporter of natural rubber, with rubber and tin providing the bulk of colonial tax revenues.

Introduction of rubber

The British also improved public health facilities, reducing the incidence of some tropical diseases, and they facilitated the establishment of government Malay and Christian mission (mostly English-language) schools; the Chinese generally had to develop their own schools. These separate school systems, however, helped perpetuate the pluralistic society. Some Chinese, Malays, and Indians benefited from British economic policies, while others enjoyed no improvement or saw living standards drop. Government-sanctioned opium and alcohol use provided a major revenue source in some areas.

Between 1800 and 1941 several million Chinese entered Malaya (especially the west-coast states), Sarawak, and British North Borneo to work as labourers, miners,

planters, and merchants. South Indian Tamils were imported as the workforce in Malayan rubber estates. Malays accounted for 90 percent of Malaya's population in 1800, but by 1911 they constituted only about 60 percent. A pluralistic society was developing, with most Malays in villages, Chinese in towns, and Indians on plantations. Colonial authorities skillfully utilized "divide and rule" tactics to maintain their control. Through enterprise, organization, and cooperation, many Chinese became part of a prosperous, urban middle class that controlled retail trade. The various ethnic groups generally lived in their own neighbourhoods, followed different occupations, practiced their own religions, spoke their own languages, operated their own schools, and later formed their own political organizations. Some mostly ethnically oriented nationalist currents stirred in Malaya, Singapore, and Sarawak by the 1930s. Malay groups either pursued Islāmic revitalization and reform or debated the future of the Malays in a plural society, while Chinese organizations were oriented toward political trends in China.

The Borneo states experienced many of the same changes. Sir Charles Brooke, who governed Sarawak from 1868 to 1917, succeeded his uncle, passing the state on to his son, Charles Vyner de Windt Brooke (ruled 1917–46); they furthered the Brooke pattern of personal rule. Chinese immigrants came in response to economic incentives, and by 1939 they accounted for a quarter of the state population. The Brookes involved the Malay elite in government. Sarawak society came to be characterized by a three-way division: most Malays in government or fishing; most Chinese in trade, labour, and cash-crop farming; and most Iban in the police force or shifting cultivation. Gambier and pepper were planted, with Sarawak becoming the major world supplier of the latter crop. Later, rubber became dominant, and an oil industry developed. Most cash-crop agriculture remained in smallholdings rather than following the plantation pattern characteristic elsewhere. Christian missionary activity and church, Chinese, and Malay schools generated sociocultural change. In the 1930s both the Chinese and Malay communities experienced rising ethnic consciousness as personal rule began to erode.

The North Borneo Company in Sabah contrasted with the Brookes. It concentrated on developing an extractive economy for the benefit of its shareholders, based mostly on Western-owned tobacco and rubber estates and forest exploitation. Christian missions facilitated change among non-Muslims. Immigrant Chinese and Indonesians provided a plantation workforce. Both the Brookes and the company created single states out of many local societies, but they tolerated little open political activity.

Japanese occupation of Malaya

Political transformation. The occupation of Malaya and Borneo by Japan (1942–45) during World War II generated tremendous changes in those territories. Their economies were disrupted, and communal tensions were exacerbated because Malays and Chinese reacted differently to Japanese control. The Japanese desperately needed access to the natural resources of Southeast Asia; they invaded Malaya in December 1941, having neutralized American military power in Hawaii (Pearl Harbor) and the Philippines, and shortly controlled the peninsula, Singapore, and Borneo. Pro-communist, predominantly Chinese guerrillas waged resistance in Malaya, and a brief Chinese-led revolt also erupted in North Borneo. In many places increasing politicization and conflict within and among ethnic groups developed as a result of economic hardship and selective repression; the rule of the Brookes and of the North Borneo Company was permanently undermined, while in Malaya some Chinese and Malays saw that British domination was not inevitable. Nonetheless, most people welcomed the Japanese defeat in 1945.

After the end of the war, Sarawak and North Borneo became British crown colonies, but Sarawak faced a turbulent political situation. Many Malays opposed the termination of Brooke rule and Sarawak's cession to Britain; the resulting sociopolitical divisions persisted for years. With the establishment of the British North Borneo colony, the capital was moved to Jesselton (now Kota Kinabalu). Some local self-government was introduced in Malaya. The major generator of political organizing, however, was a British proposal to form a single Malayan Union, incorporating all the Malayan territories except Singapore, that would diminish state autonomy and accord equal political and citizenship rights to non-Malays. A tremendous upsurge of Malay political feeling against this plan, led by Dato Onn bin Ja'afar, resulted in the creation in 1946 of the United Malays National Organization (UMNO) as a vehicle for Malay nationalism and political assertiveness. Strikes, demonstrations, and boycotts doomed the scheme, and the British began to negotiate with the UMNO about the Malayan future.

The negotiations resulted in the creation of the Federation of Malaya in 1948, which unified the territories but provided special guarantees of Malay rights, including the position of the sultans. These developments alarmed the more radical and impoverished sectors of the Chinese community. In 1948 the Communist Party of Malaya—a mostly Chinese movement formed in 1930 that had provided the backbone of the anti-Japanese resistance—went into the jungles and began a guerrilla insurgency to defeat the colonial government, sparking a 12-year period of unrest known as the Emergency. The communists waged a violent and ultimately unsuccessful struggle supported by only a minority of the Chinese community. The British struggled to suppress the insurgency by military means, including an unpopular strategy that forcibly moved many rural Chinese into tightly controlled New Villages. Although this policy isolated villagers from guerrillas, it also increased the government's unpopularity. The British finally achieved success when, under the leadership of British high commissioner Sir Gerald Templer, they began addressing political and economic grievances as well, increasingly isolating the rebels. Promising independence, British officials began negotiating with the various ethnic leaders, including the UMNO and the Malayan Chinese Association (MCA), formed in 1949 by wealthy Chinese businessmen. A coalition consisting of the UMNO (led by the aristocratic moderate Tunku Abdul Rahman), the MCA, and the Malayan Indian Congress contested the national legislative elections held in 1955 and won all but one seat; this established a permanent political pattern of a ruling coalition—known first as the Alliance Party and later as the National Front—that united ethnically based, mostly elite-led parties of moderate to conservative political leanings, with the UMNO as the major force.

The Emergency

On Aug. 31, 1957, the Federation of Malaya achieved independence (*merdeka*) under an Alliance government headed by Tunku Abdul Rahman as prime minister. Singapore, with its predominantly Chinese population, remained outside the federation as a British crown colony. The arrangement tended to favour the Malays politically, with UMNO leaders holding most federal and state offices and the kingship (*yang di-pertuan agong*) rotating among the various Malay sultans, but the Chinese were granted liberal citizenship rights and maintained strong economic power. Kuala Lumpur became the federal capital.

New currents also were emerging in Borneo. Colonial rule succeeded in rebuilding and expanding the economies of the two colonies, with rubber and timber providing the basis for postwar economic growth. Health and education facilities only slowly permeated outside the towns. Political consciousness began to spread, however, as elections were held for local councils. During the 1950s the development of radio broadcasting and newspapers particularly stimulated the Kadazan community to become involved in Sabah politics, while, in Sarawak, Chinese and Malay leaders formed the first political parties there—some espousing multiethnic identities—in expectation of independence. Political activity accelerated with the mooting of the proposal for a federated Malaysian state by Malayan and British officials in 1961, and new parties formed in Sabah representing the Kadazan, Chinese, and Muslim communities. Statewide elections were held in Sabah and Sarawak, with most of the parties accepting independence through merger with Malaysia; that sentiment increased after the Philippines claimed Sabah, based on former Sulu suzerainty.

British leaders proposed a Malaysian federation as a way of terminating their now burdensome colonial rule over

Singapore, Sarawak, and Sabah, even though these states were historically and ethnically distinct from Malaya and from each other. It was in many ways to be a marriage of convenience. Malaya was closely linked economically with bustling Singapore, and the Malays felt a kinship to the various Muslim groups in Borneo. Tunku Abdul Rahman believed the federation could defuse potential leftist Chinese activity while balancing the Chinese majority in Singapore with the non-Chinese majorities of the Borneo states. Malaya already contained a Chinese minority of nearly 40 percent, with Malays barely in the majority there.

Formation of the federation

Hence, on Sept. 16, 1963, the Federation of Malaysia was formed, with Sarawak and Sabah (East Malaysia) shifting from a Bornean to a peninsular orientation. Brunei, which had been invited to join, chose to remain a British protectorate and later became independent as a small, oil-rich Malay sultanate.

Malaysia. The new, hurriedly formed nation faced many political problems, including a period of Indonesian military opposition that ended in 1966, sporadic communist insurgency in Sarawak, periodic East Malaysian disenchantment over Malayan domination and federal policies, and the secession of Singapore from the Federation (at Malaysia's urging) in 1965. The latter event resulted from increasing friction between the mostly Malay federal leaders and the mostly Chinese state leaders, especially Singapore's independent-minded chief minister, Lee Kuan Yew, who disagreed on national goals. Under Lee's autocratic direction and freewheeling economic policies, Singapore became a highly prosperous but tightly controlled city-state, and relations with Malaysia gradually improved. Both countries became founding members of the Association of Southeast Asian Nations (ASEAN) in 1967.

The secession of Singapore allowed the UMNO to exercise more influence over federal policies, even if it did not end political uncertainties. Communal tensions on the peninsula following a heated election generated riots and a nationwide state of emergency in 1969–70. Many non-Malays resented the government's attempts to build national unity and identity, such as increasing the use of the Malay language in education and public life. Government policies aimed at redistributing more wealth to Malays, as well as a growing Islāmic revival, particularly worried the Chinese. The New Economic Policy, launched in 1971 and renewed as the New Development Policy in 1991, was designed to greatly increase Malay wealth and economic potential. Beginning in the late 1970s, the Islāmic fundamentalist revival, or *dakwah* movement, increasingly attracted the support of young Malays who had become alienated by the growth of a Westernized, materialistic society, and this generated divisions within Malay society. Rural development policies reduced poverty rates, but large pockets of urban and especially rural poverty persisted, with many regional and ethnic inequities in the distribution of wealth. Radical critics of the government (including communists, socialists, Islāmic militants, and progressive intellectuals) were politically marginalized or sometimes detained.

For Sarawak and Sabah, politics within Malaysia has proved a turbulent experience. The decision to join was made in haste, and many people continued to resent the loss of their autonomy, especially control over growing petroleum revenues. Political crises have occurred periodically in Sarawak, although it has been governed since 1970 by a Malay-dominated, pro-federal but multiethnic coalition that represented a triumph of peninsular alliance-style politics. By the mid-1980s, however, some Iban leaders had challenged the coalition for being too accommodating to wealthy Malay and Chinese interests. The government encouraged the assimilation of Sarawak society with that of the peninsula and dramatically increased the exploitation of timber resources, sometimes at the expense of powerless interior peoples. Sabah politics also have proved to be contentious, with chronic tensions between Muslim and non-Muslim groups. Between 1967 and 1975 Chief Minister Tun Mustapha ruled the state with an iron hand, co-opting or repressing opponents, promoting Islām, and challenging federal policies. The multiethnic coalition that replaced Mustapha continued to preside over rapid economic growth purchased by the exploitation of Sabah's bountiful natural resources. Communal tensions surfaced again in 1987, when a Christian Kadazan-led party swept into power and followed policies opposed by federal leaders. Although peninsular sociopolitical patterns increasingly influenced Sabah and Sarawak, the states remained unique within the Malaysian system.

Politics in Sabah

Since 1963 Malaysia has maintained a quasi-democratic parliamentary political system that includes regular elections and moderate political diversity but also some restrictions on civil liberties, including a ban on public discussion of "sensitive" issues. Tunku Abdul Rahman was succeeded as prime minister by Tun Abdul Razak in 1970. On Abdul Razak's death in 1976 another UMNO leader, Datuk Hussein Onn, replaced him. In 1981 Mahathir bin Muhammed became prime minister, the first nonaristocrat to hold that office. Mahathir's assertive style and controversial policies generated a major split in the UMNO; in 1988 Mahathir outmaneuvered his opponents, dissolving the UMNO and forming a new Malay party, UMNO Baru (New UMNO). Government and business leaders have managed to develop a prosperous, diversified economy, although commodity exports have remained important and certain areas have experienced severe environmental problems. Malaysia's literacy rates have risen dramatically, and the government has constructed an extensive public education system. The large and expanding urban middle class has become increasingly multiethnic, with a growing percentage of non-Malays fluent in the national language. Although development policies have been criticized as lacking ethnic and regional balance, Malaysia nonetheless has achieved considerable success at creating national unity and sociopolitical stability out of deep regional and ethnic divisions. (C.A.Lo.)

For later developments in the history of Malaysia, see the BRITANNICA BOOK OF THE YEAR.

Myanmar

Myanmar is an independent country in the western portion of mainland Southeast Asia, with an area of 261,228 square miles (676,577 square kilometres). It is bordered by China to the north and northeast, Laos to the east, Thailand to the southeast, the Andaman Sea and Bay of Bengal to the south and southwest, Bangladesh to the west, and India to the northwest.

In 1989 the country's official English name was changed from the Union of Burma to the Union of Myanmar (Burmese: Pyidaungzu Myanma Naingngandaw); in the Burmese language the country has been known as Myanma (or, more precisely, Mranma Prañ) since the 13th century. Also in 1989, the English name of the capital, Rangoon, was dropped in favour of the common Burmese name, Yangôn. In this discussion, the name Burma is used for the country during the period of British rule (1885–1948) and during the subsequent period of independence until 1989; the name Myanmar is used in all other contexts.

Myanmar stretches from latitude 10° N to about 28° 30′ N. It is thus the northernmost of the Southeast Asian countries, with considerable territory situated outside the tropics. The country is shaped like a kite with a long tail that runs south along the Malay Peninsula. Its total length from north to south is about 1,275 miles (2,050 kilometres), and its width at the widest part, across the centre of the country at about the latitude of Mandalay, is approximately 580 miles from east to west.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Myanmar slopes from north to south, from an elevation of 19,296 feet (5,881 metres) at Mount Hkakabo (the country's highest peak) in the extreme north to sea level at the Irrawaddy (Ayeyarwady) and Sittang (Sittoung) river deltas. The mountain ranges generally run from north to south. The country as a whole can be divided into five physiographic regions—the northern mountains, the western ranges, the eastern plateau, the central basin and lowlands, and the coastal plains.

Five physiographic regions

The northern mountains consist of a series of ranges that form a complex knot at Mount Hkakabo. Geologically,

this knot marks the northeastern limit of the encroaching Indian-Australian Plate, which has been colliding with the southern edge of the Eurasian Plate for roughly the past 50 million years and thrusting up the mountain ranges of Myanmar and beyond. This region contains the sources of several of Asia's great rivers: the Irrawaddy, which rises and flows wholly within Myanmar, and the Salween (Thanlwin), Mekong, and Yangtze, which rise to the north in China. The upper courses of these rivers all flow through deep gorges within a short distance of each other, separated by steep, sheer peaks.

The western ranges traverse the entire western side of Myanmar from the northern mountains to the southern tip of the Arakan (Rakhine) Peninsula, where they run under the sea and reappear as the Andaman and Nicobar islands. Their average height is about 6,000 feet, although some peaks rise to 10,000 feet and higher. The mountains consist of old crystalline rocks surrounded by hard, tightly folded sedimentary rocks on either side. From north to south, the Pātkai Range, Nāga Hills, and Chin Hills form the border between India and Myanmar. To the south of these are the Arakan Mountains, which lie entirely within Myanmar and separate the coastal strip from the central basin.

The Shan Plateau in the east rises abruptly from the central basin, often in a single step of 2,000 feet. Occupying the eastern half of the country, it is deeply dissected, with an average height of 3,000 feet. The plateau was formed during the Mesozoic Era (245 to 66.4 million years ago) and thus is a much older feature than the western mountains, but the plateau also shows more recent and intensive folding, with north-south longitudinal ranges reaching elevations of 6,000 to 8,600 feet rising abruptly from the plateau surface. Northward, the plateau merges into the northern mountains, and southward it continues into the Dawna Range and the peninsular Tenasserim Mountains, each a series of parallel ranges with narrow valleys.

The central basin and lowlands, lying between the Arakan Mountains and the Shan Plateau, are structurally connected with the folding of the western ranges. The basin was deeply excavated by the predecessors of the Irrawaddy, Chindwin (Chindwinn), and Sittang rivers; the ancient valleys are now occupied by these rivers, which cover the ancient soft sandstones, shales, and clays with their more recent alluvial deposits. In the deltaic regions formed by the Irrawaddy and Sittang rivers, the landscape is absolutely flat, and the monotony is relieved by only a few blocks of erosion-resistant rocks that are never more than 60 feet high. The basin is divided into two unequal parts, the larger Irrawaddy valley and the smaller Sittang valley, by the Pegu Mountains. In the centre of the basin and structurally connected with the Pegu Mountains and its northern extension is a line of extinct volcanoes with small crater lakes and eroded cones, the largest being Mount Popa, at 4,981 feet.

The coastal areas consist of the narrow Arakan and Tenasserim coastal plains, which are backed by the high ranges of the Arakan and Tenasserim mountains and are fringed with numerous islands of varying sizes.

Drainage and soils. Like the mountains, Myanmar's main rivers run from north to south. About three-fifths of Myanmar's surface is drained by the Irrawaddy and its tributaries. Flowing entirely through Myanmar, it is navigable for nearly 1,000 miles. At the apex of its delta, the Irrawaddy breaks up into a vast network of streams and empties into the Andaman Sea through nine mouths. Its great tributary, the Chindwin, drains the western region. The Bassein River drains the southern Arakan Mountains, and the Yangôn (Rangoon) River drains the Pegu Mountains, both entering the Irrawaddy at the delta. The Sittang flows into the Gulf of Martaban of the Andaman Sea and, in spite of its comparative shortness, has a relatively large valley and delta. The Shan Plateau is drained by the Salween River, which enters Myanmar from southern China and empties into the Gulf of Martaban southeast of the Sittang. It is deeply entrenched and crosses the plateau in a series of deep gorges. Many of its tributaries are more than 300 miles long and join the Salween in cascades. The Arakan coastal plains are drained by short, rapid streams, which, after forming broad deltas, flow into the Bay of Bengal. The Tenasserim plains also are drained by short and rapid rivers, which enter the Gulf of Martaban.

Rivers

The highland regions of Myanmar are covered with highly leached dark red and reddish brown latosols. When protected by forest cover, these soils absorb the region's heavy rain, but they erode quickly once the forest is cleared. The lowland regions are covered with alluvial soils—mainly silt and clay. Low in nutrients and organic matter, they are improved by fertilizers. In the central-region dry belt are found red-brown soils rich in calcium and magnesium. In the same region, however, when the soil has a low clay content, it becomes saline under high evaporation and is recognizable by its yellow or brown colour.

Climate. Although Myanmar is located in the monsoon region of Asia, its climate is greatly modified by its geographic position and its relief. The cold air masses of Central Asia bring snow to the northern mountains for two months of the year, but this mountain wall prevents the cold air from moving farther south, so that Myanmar lies primarily under the influence of the monsoon winds. The north-south alignment of ranges and valleys creates a pattern of alternate zones of heavy and scanty rainfall during both the northeast and southwest monsoons. Most of the precipitation, however, comes from the southwest monsoon. The west coast is subject to occasional tropical storms called cyclones.

Myanmar has three seasons: the cool, relatively dry northeast monsoon (late October to mid-February), the hot, dry intermonsoonal season (mid-February to mid-May), and the rainy southwest monsoon (late May to late October). The coastal regions and the western and southeastern ranges receive more than 200 inches (5,100 millimetres) of precipitation annually, while the delta regions receive about 100 inches. The central region is not only away from the sea but also in the rain shadow of the Arakan Mountains. Rainfall gradually decreases northward until in the dry zone it is only between 20 and 40 inches. The Shan Plateau, because of its elevation, receives between 75 and 80 inches.

Elevation and distance from the sea affect temperature as well. Although Myanmar generally is a tropical country, temperatures are not uniformly high throughout the year. The daily temperature range is greater than in nearly all other parts of Southeast Asia, but no locality has a continental type of climate (*i.e.*, one characterized by large seasonal differences in average temperature). Mandalay, in the centre of the dry zone, has some of the greatest daily temperature ranges, which average about 22° F (12° C) annually. The average daily temperature at Mandalay is 82° F (28° C), compared to 81° F (27° C) at Yangôn near the coast, 79° F (26° C) at Sittwe (Akyab) in Arakan (Rakhine) state, and 71° F (22° C) at Lashio on the Shan Plateau.

Temperatures

Plant and animal life. Even after centuries of rice cultivation involving clearing large areas of forest, about half of Myanmar is covered with forests of various types, depending on elevation and the amount of precipitation. Subtropical and temperate forests of oak and pine are found at elevations above 3,000 feet. In the northern mountains, above 6,000 feet, are forests of rhododendrons. Tropical, evergreen rain forests of hardwood trees occur in areas receiving more than 80 inches of rain annually. In regions where the rainfall is between 40 and 80 inches are found broad-leaved, tropical-deciduous "monsoon" forests, the trees of which shed their leaves during the hot season. They produce valuable woods, notably teak. Where rainfall is less than 40 inches, the forests gradually open into scrubland. There are no true grasslands in Myanmar, but bamboo, bracken, and coarse grass grow in areas where the forest has been cleared and then abandoned. In the Irrawaddy and Sittang deltas are found tidal forests of mangrove trees that grow as high as 100 feet and supply firewood and bark for tanning.

The jungles of Myanmar are home to a profusion of birdlife, including pheasants, parrots, peafowl and other wild fowl, and grouse. The Asiatic two-horned rhinoceros, wild buffalo, gaur (wild bison), and various kinds of deer were once plentiful but are now reduced in number and

protected. Elephants are numerous, and many are trained for work. Tigers, leopards, and wildcats are still common. Bears are found in hilly regions, and gibbons and monkeys of various kinds inhabit the thicker parts of the forests. Snakes include pythons, cobras, and vipers, and crocodiles are found in the deltas. Turtles live in coastal regions, and edible fishes abound in every stream.

Traditional regions. Speakers of Burmese and Mon historically have lived in the plains, while speakers of a dialect of Burmese retaining archaic features occupied the Arakan and Tenasserim coastal plains. The hills were inhabited by those speaking Shan, Kachin, Chin, and numerous other languages. In the plains the division between northern and southern Myanmar (Upper and Lower Burma, respectively) dates from early history, not only because of differences in the geography but also because the Mon (now a small minority) lived in southern Myanmar. The northern dry zone, where the majority Burman population lived, was the cultural, political, and economic heartland of Myanmar. The division became more marked during the period 1852–85, when southern Myanmar became British Burma.

Villages

Settlement patterns. Myanmar is a land of villages. Except for a few large cities—notably Yangôn, Mandalay, and Moulmein (Mawlamyine)—the towns essentially are large villages. The hill peoples, although practicing shifting agriculture, have settled in upland villages at some distance from the fields. On the Shan Plateau and in the neighbouring river valleys, the fields adjoin the villages. Older villages are circular in shape, but along the banks of the delta streams and along railways the villages are rectangular. Houses are built of timber and bamboo, the roofs being thatched or tiled. In the past, houses typically were built on piles, the original purpose being protection from wild animals or floods. The style persists in many villages, especially those on the hills, and farm animals are kept under the houses at night. In small towns the piles have given way to a supporting brick structure with concrete flooring, the upper story still being made of timber. Houses entirely of brick were few in number before 1942, but many later sprang up in Yangôn, Mandalay, and larger towns on the rubble of buildings destroyed during World War II. Life in villages is still communal because of custom, the influence of Buddhism, and the "redistributive" and reciprocal nature of agrarian society.

The people. *Linguistic groups.* Several indigenous languages—as distinct from mere dialects—are spoken in Myanmar. The official language is Burmese, spoken by the people of both the plains and the hills. These languages belong to three language families. The Burmese language itself, and most of the other languages, belong to the Tibeto-Burman subfamily of the Sino-Tibetan family. The Shan language belongs to the Tai family. Languages spoken by the Mon of southern Myanmar and by the Wa and Palaung of the Shan Plateau are members of the Mon-Khmer subfamily of the Austro-Asiatic family.

Until colonial times, only Pyu, Burmese, Mon, and Shan were written; writing systems for Karen, Kachin, and Chin were developed later. The Burmese spoken in Arakan state and Tenasserim (Taninthary) division suggests that it has preserved the language's ancient pronunciations. For the majority of the hill peoples, Burmese is a second language.

During the colonial period, English became the official language, but Burmese continued as the primary language in all other settings. Both English and Burmese were made compulsory subjects in schools and colleges. Since a knowledge of English became an asset, many learned to speak it, and a small English-speaking elite emerged. Burmese, Chinese, and Hindi were the languages of commerce. After independence, English ceased to be the official language and, after the military coup of 1962, lost its importance in schools and colleges; an elementary knowledge of English, however, is still required, and its instruction is again being encouraged.

Ethnic groups. The original home of the Burmans in the dry zone established the ethnic character of the entire Irrawaddy valley and the coastal strips. These areas hold the majority of the population.

The Irrawaddy and Sittang deltas were once peopled by the Mon, who may have entered the country from their kingdoms in the Chao Phraya River valley in Thailand. They were conquered in the 11th century by the Burmans, a more martial and less cultured group at the time. The Mon attempted twice to throw off Burman control, but by the end of the 18th century they had been largely absorbed by the Burmans, by intermarriage as well as by suppression. A sizable number still remain in the Sittang valley and in Tenasserim; although they continue to call themselves Mon, most of them have been integrated into Burman culture and no longer speak their original language.

Tribal peoples

In the western hills and the Chindwin River valley are various groups called by the comprehensive name of Chin. The upper Irrawaddy valley and the northern hills are occupied by groups under the comprehensive name of Kachin. These peoples have had a long association with the Burmans.

The Shan of the Shan Plateau have little ethno-linguistic affinity with the Burmans, and their society, unlike that of the plains peoples, was less elaborately structured. The Wa and the Palaung are Mon-Khmer speakers, but, because of the smallness of their numbers and their long residency on the plateau, they are sometimes confused with the Shan. In the same way, the Nāga on the Myanmar side of the frontier with India are mistakenly placed with the Chin, and the Lolo-Muhso in northeastern Myanmar are grouped with the Kachin.

The Karen are the only hill people who have settled in significant numbers in the plains. Although ethnically and linguistically Tibeto-Burman, they share territory and much vocabulary with the Mon. They are found in the deltas among the Burmans, in the Pegu Mountains, and along both sides of the lower Salween River. The Kayah, who live on the southern edge of the Shan Plateau, were known as Red Karens, or Karenni, apparently from their red robes. Although ethnically and linguistically Karen, they tend to have their own identity.

During the period of British colonial rule, there were sizable communities of South Asians and Chinese, but many of these people left at the outbreak of World War II. A second exodus took place in 1963, when commerce and industry were nationalized.

Religious groups. The vast majority of the population is Theravāda Buddhist. The Burmans are Buddhists except for minimal numbers of Christians and Muslims; the Shan also are Buddhists. Among the Karen, there are many more Buddhists than Christians. The other hill peoples are animists except for a small number of Kayah Buddhists and Kachin and Chin Christians, and even they practice animism to some degree.

Demographic trends. The population density in each region is related to agricultural production, particularly of rice. Thus, the most populous regions are the Irrawaddy delta and the dry zone, with the highest densities found in the upper delta, between Yangôn and Henzada. The populations of the Sittang delta, the sedimented hinterland of Sittwe, and the regions of both sides of the lower Chindwin River are moderately dense. Arakan (except the Sittwe region), the west bank of the Irrawaddy at the base of the Arakan Mountains, Tenasserim, and the more inaccessible parts of the western and northern mountains and the Shan Plateau are sparsely inhabited.

The economy. Myanmar's economy is one of the least developed of the region and is basically agricultural; more than two-thirds of the people derive their livelihoods directly from agricultural pursuits. Of the nonagricultural workers who are employed in the other sectors of the economy, many are indirectly involved in agriculture through such activities as transporting, processing, marketing, and exporting agricultural goods.

Government policies

Nearly half of Myanmar's economic output—notably all large industrial enterprises, the banking system, insurance, foreign trade, domestic wholesale trade, and nearly all the retail trade—was nationalized in 1962–63. Small-scale industry (consisting mainly of food and beverage processing, miscellaneous manufacturing, and cottage industries), agriculture, and fishing were left in the private sector. In 1975–76, however, the government placed nationalized

corporations on a commercial basis and instituted a bonus system for workers. The overall economic objectives of self-sufficiency and the exclusion of foreign investment also were revised. Foreign investment was permitted to resume in 1973 and was further liberalized in the late 1980s.

Enterprises remaining in the private sector after nationalization account for only a small fraction of the nation's tax income. The balance is collected from the public sector. The principal sources of revenue are taxes (income, commercial, and customs) and receipts from state enterprises.

Myanmar also has an informal economy. Considerable quantities of consumer goods are smuggled into the country, and teak and gems are exported both legally and illegally. In addition, northern Myanmar is one of the largest producers of opium in the region.

Agriculture. Myanmar may be divided into three agricultural regions: the delta, where rice cultivation predominates; the dry zone, an area largely of rice production but also where a wide variety of other crops are raised; and the hill and plateau regions, where forestry and shifting agriculture are the most important.

Although the dry zone was Myanmar's most important agricultural region in the past, the rice production of the Irrawaddy delta now provides much of the country's export earnings and the staple diet of the country's people. About half of all agricultural land in Myanmar is devoted to rice, and, despite a climate that permits much more extensive double-cropping, only a small proportion of the land is actually so managed. The delta's traditional agriculture consisted primarily of rice in normal years, with the substitution of millet in drier years when there was insufficient moisture for rice; both grains yielded good returns on the alluvial soils. After Burma was officially annexed to British India in 1886, however, colonial policy called for a more commercially oriented and extensive cultivation of rice. Since the indigenous labour force was thought to be insufficient to support the colonial export economy, the immigration of Indian and Chinese labourers was officially encouraged during the early decades of the 20th century. By 1942, first-generation immigrants made up about 13 percent of the total population. Despite relatively low growth in rice production after World War II, rice remained both the basic food and the basic export of Myanmar.

Crops raised in the dry zone, in addition to rice, include wheat, millet, corn (maize), peanuts (groundnuts), sesame, legumes, tea, and rubber. To cultivate much of this land successfully, however, irrigation is required. The earliest known irrigation works were constructed in the 1st century and greatly improved in the 11th century; though their maintenance has lapsed somewhat since the fall of the monarchy, many are still in active service. As in the delta, the arrival of the British in the dry zone led to increased commercial and public-works activities. British authorities repaired and extended parts of these ancient systems during the early 20th century. Most of Myanmar's irrigated land is in the dry zone, and almost all of it is planted in rice. The portions of the dry zone that are not irrigated are utilized for the production of crops that are less sensitive to the seasonality or irregularity of rainfall than rice. In addition to the crops mentioned above, cotton and sugarcane are cultivated, although neither is of considerable significance. Cattle also are raised there.

Irrigation

The third agricultural zone, the hill and plateau country, occupies perhaps two-thirds of the area of Myanmar. Although this land has less economic significance than the other two zones, it is the home of many of the country's non-Burman ethnic groups. They generally continue to practice shifting cultivation (called *taungya* in Burmese), although more sedentary modes also exist and others are imposed with the advance of agricultural technology and central planning. Outside the forest areas of these highlands, the principal crops raised are rice, yams, and millet, and large numbers of pigs and poultry are kept. Bullocks and buffalo are used as beasts of burden, and goats, pigs, and poultry are raised for food in all parts of the country.

Fisheries and forestry. The second most important element in the diet after rice is fish—fresh or in the form of *ngapi,* a sort of paste that is prepared in a variety of ways and eaten as a condiment. Marine fisheries are not well-developed, although the reported commercial catch is more than three times as great as the reported catch from inland waters. Much private, noncommercial fishing is provided, however, in virtually every type of permanent, seasonal, or artificial body of inland water of any size. Two nonindigenous fish, the European carp and the tilapia (originally brought from Thailand), have been introduced and have bred well in impounded waters.

Forestry has been particularly important as a source of foreign exchange. Myanmar is estimated to have the bulk of the world's exploitable teak supplies. Teak is found in the tropical-deciduous forests of the hills. The forests are owned and regulated by the state, but concern has been raised about illegal logging.

Mining and power. The modern development of Myanmar's rich mineral deposits began in the mid-1970s. Deposits of silver, lead, and zinc are worked in the northern Shan Plateau, tin and tungsten in Tenasserim, and barite from the Maymyo area. Copper mining at Monywa began in the early 1980s. Rubies and sapphires have been mined in the northern Shan Plateau since precolonial times. Jade is mined in the northern mountains. Oil and natural gas are produced for domestic consumption. Coal is found in the upper Chindwin valley.

The demand for electricity chronically has outstripped capacity. The government has built several hydroelectric power plants, including those on the Balu River (a tributary of the Salween), at Taikkyi near Pegu (Bago), in northern Arakan state, and near Mandalay. Hydroelectricity now accounts for nearly half of Myanmar's total generating capacity.

Hydroelectricity

Industry. There was little industrialization until after independence, when a limited program began. Yangôn, Myingyan (in the dry zone), and Arakan state were selected to become the new industrial centres. There are textile factories at Yangôn and Myingyan and one near Paleik in the central region. Oil refineries are located at Chauk, Syriam, and Mann. Yangôn also has steel-processing and pharmaceutical plants, and there is a paper mill in Arakan. Existing food-processing plants (mainly rice mills) and lumber mills have been improved and expanded. Cottage industries are encouraged by subsidies.

Trade. The decision in the early 1960s to limit foreign trade reversed the export orientation of the British colonial period. Subsequent relaxation of trade restrictions, especially in the late 1980s, has again allowed trade to become a significant component of the national economy. Myanmar's economy remains dependent on the export of commodities, mainly rice, teak, and minerals and gems. It imports machinery and equipment, industrial raw materials, and consumer goods. Myanmar's chief trading partners are Japan, the European Community, other Southeast Asian nations (especially Singapore), India, and China.

Transportation. The Irrawaddy River is the backbone of Myanmar's transportation system. Trade in rice is dependent on water transport. The Irrawaddy is navigable year-round up to Bhamo and to Myitkyina during the dry season when there are no rapids. The Chindwin is navigable for some 500 miles from its confluence with the Irrawaddy below Mandalay. The many streams of the Irrawaddy delta are navigable, and there is a system of connecting canals. The Sittang, in spite of its silt, is usable by smaller boats; but the Salween, because of its rapids, is navigable for less than 100 miles from the sea. Small steamers and country boats also serve the coasts of Arakan and Tenasserim.

The first railway line, running from Rangoon (Yangôn) to Pye (Prome) and built in 1877, followed the Irrawaddy valley. The line was not extended to Mandalay; instead, after 1886 a new railway from Rangoon up the Sittang valley was constructed, meeting the Irrawaddy at Mandalay. From Mandalay it crosses the river and, avoiding the Irrawaddy valley, goes up the Mu River valley to connect with the Irrawaddy again at Myitkyina. A short branch line connects Naba to Katha on the Irrawaddy below Bhamo.

The Yangôn-Mandalay-Myitkyina railway is the main artery, and from it there are branch lines connecting the

Railways

northern and central Shan Plateau with the Irrawaddy. Other branches run from Pyinmana across the Pegu Mountains to Kyaukpadaung and from Pegu to Moulmein to Ye. The Pye-Yangôn railway has a branch line crossing the apex of the delta to Henzada and Bassein (Pathein).

The road system, until independence, was confined to the Irrawaddy and Sittang valleys, duplicating the railway route. A road goes from Pye along the Irrawaddy to the oil fields. Government policy is to improve and extend existing highways and to construct new ones. There were originally three international roads in use during World War II—the Burma Road from Lashio to K'unming in China; the Stillwell, or Ledo, Road between Myitkyina and Ledo in India; and the road between Kêng Tung in the southeastern Shan Plateau and northern Thailand. These roads subsequently became neglected but more recently were rebuilt and extended.

The state-run airline runs frequent domestic flights between Yangôn and other cities, and Yangôn has an international airport. Yangôn, as the terminus of road, rail, and river transport systems, is the country's major port, with up-to-date equipment and facilities. Bassein, Moulmein, and Sittwe are also important ports.

Administration and social conditions. *Government.* Myanmar's constitution came into force on Jan. 4, 1974, the 26th anniversary of the country's independence. According to the constitution, supreme power rested with a unicameral People's Assembly (Pyithu Hluttaw), which exercised legislative, executive, and judicial authority. By the late 1980s the assembly consisted of 489 members, who were elected to four-year terms.

The organs of the People's Assembly were the Council of State, the Council of Ministers (Cabinet), the Council of People's Justices, the Council of People's Attorneys, and the Council of People's Inspectors. The Council of State consisted of 29 members: one representative elected from each of the country's 14 states and divisions, an equal number elected by the People's Assembly as a whole, and the prime minister as an ex officio member. The Council of State elected its own chairman, who was ex officio president of the country, and its own secretary. The president and the secretary were also, respectively, the chairman and the secretary general of the Burma Socialist Programme Party (BSPP). The Council of State appointed senior civil servants and deputy ministers and submitted lists of names for election by the assembly of the prime minister and the councils of ministers, justices, attorneys, and inspectors. The Council of People's Justices was the equivalent of a supreme court in a parliamentary democracy; the Council of People's Attorneys was comparable to an attorney general and the Council of People's Inspectors to an auditor general.

From 1962 to 1988 the official political party was the BSPP. Membership at first was restricted to cadres, but, after the party held its first national assembly in 1971, its membership was widened so that it became a national party. Civil servants and members of the armed forces, as well as workers and peasants, were members, and senior military officials and civil servants were included in the party's hierarchy.

In September 1988 the armed forces took control of the government, creating a new ruling body, the State Law and Order Restoration Council (SLORC). All state organs, including the People's Assembly, the Council of State, and the Council of Ministers, were abolished, and their duties were assumed by the SLORC. A Supreme Court was established as the supreme judicial authority, its members appointed by the SLORC.

The SLORC

The law maintaining the BSPP as the sole political party was abrogated, and new parties were encouraged to register for general elections to a new Constituent Assembly, which would revise the 1974 constitution. More than 90 parties participated in the elections, which were held in May 1990; of these the most important were the ruling BSPP, which had changed its name to the National Unity Party (NUP), and the main opposition party, the National League for Democracy (NLD). The NLD won an overwhelming majority of seats to the new assembly, but the SLORC did not permit the assembly to convene until January 1993.

Myanmar is divided administratively into seven states based largely on ethnicity—Arakan (Rakhine), Chin, Kachin, Karen (Kayin), Kayah, Mon, and Shan—and seven more truly administrative divisions of Myanmar proper—Irrawaddy (Ayeyarwady), Magwe (Magway), Mandalay, Pegu (Bago), Yangôn, Sagaing, and Tenasserim (Taninthary). Until 1988 there were several levels of local People's Councils—division or state, township, and ward or village—that followed the pattern of the People's Assembly; local and national elections were held simultaneously. Every council had an Executive Committee and a Judges' Committee, and all but the village or ward councils also had a Committee of Inspectors. The Judges' Committee sat as the local court, exercising criminal and civil jurisdiction. In 1988 the SLORC dissolved these bodies and assumed control of local administration.

Armed forces. Myanmar's armed forces consist of an army, a navy, and an air force. The army is by far the largest and best-equipped of the three branches, and for a number of years it has borne the chief responsibility for combating armed insurgency within the country. Members of the armed forces are recruited from throughout the country, and military service is a prime means of improving socioeconomic status. The police force, although armed and equipped and often used as a branch of the army in emergencies, remains essentially civilian in character and regional in organization.

Health and welfare. The BSPP government gave special attention to workers and peasants and to the hill peoples. In spite of a shortage of imported building materials, the

Frank Folwell/Tony Stone Images

Ruins of ancient Buddhist temples, Pagan, Myanmar.

housing problem was stabilized somewhat. A high mortality rate has been lowered substantially, although infant mortality rates have remained higher than in all other Southeast Asian states except Cambodia and Laos. Every village has a health unit and access to a hospital, but the lack of adequate sanitation and a shortage of medicines have contributed to relatively high rates of gastrointestinal diseases, tuberculosis, and malaria.

Literacy

Education. Myanmar has a long tradition of educational achievement. The literacy rate always has been high and has continued to improve. Education is compulsory between ages five and nine (primary school) and is free at primary, lower secondary (middle), and vocational schools; upper secondary (high) schools and universities charge nominal fees. The University of Yangôn (Rangoon) and the University of Mandalay (until 1958 a branch of the University of Yangôn) are the oldest and best-known institutions of higher education.

Cultural life. Myanmar's traditional culture is an amalgam of folk and royal culture. Buddhism has been a part of Myanmar's culture since the 1st century AD and has blended with non-Buddhist beliefs. The most conspicuous manifestation of Buddhist culture is the magnificent architecture and sculpture of Myanmar's many temples and monasteries, notably those at Pagan, Mandalay, and Yangôn.

In 1886 the traditional drama appeared to be dying with the elimination of the monarchy, but it had permeated the masses and survived as part of the folk tradition. With the growth of nationalism and regaining of independence, it gathered new strength. The most popular dramatic form is the pwe, which is performed outdoors. There are a variety of pwe genres, but most often the subject matter is taken from the Jātakas, the stories of the former lives of the Buddha. Traditional musical forms, influenced by those from neighbouring lands, are highly percussive. Dance forms are derived largely from southern India.

Wood carving, lacquerwork, goldwork, silverwork, and the sculpting of Buddhist images and mythological figures also survived during colonial rule; there has been a revival of these indigenous art traditions under government patronage. Both the arts of bronze casting among the Burmans and of making bronze drums among the Karen, however, disappeared. The traditional marionette show also declined, although occasionally there have been attempts to revive it. The cinema and rock music are two Western art forms that have been accepted in the cultural life of Myanmar.

Burmese literature is an intimate blend of religious and secular genres. It remained alive throughout the colonial period and, both in verse and prose, has continued to thrive. A later (though not entirely new) development was biography, which has become more popular than fiction. Government-sponsored awards are given annually for the best translation, the best novel, and the best biography.

There are state schools of dance, music, drama, and fine arts at Yangôn and Mandalay. The National Museum is at Yangôn, and there are regional museums at Pagan, Mandalay, and other regional centres.

(M.H.Au./M.A.A.-T.)

For statistical data on the land and people of Myanmar, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

In prehistoric times Myanmar was inhabited along its coasts and its river valleys. During most of the 1st millennium AD the overland trade route between China and India passed through Myanmar's borders, and merchant ships from India, Sri Lanka, and even farther west converged on its ports, some of which also were the termini of the portage routes from the Gulf of Thailand across the narrow Isthmus of Kra on the Malay Peninsula. Thus, Myanmar often was the western gateway of mainland Southeast Asia.

Indian influences

The Indian merchants brought with them not only precious cargoes but their religious, political, and legal ideas; and within a few decades Indian cultural traditions had remolded indigenous society, thought, and arts and crafts. Yet important components of Myanmar's own native culture were retained, creating a lasting synthesis with Indian culture. Surrounded on three sides by mountains and on the fourth by the sea, Myanmar always has been somewhat isolated; as a consequence, its culture has remained distinct in spite of the many Indian influences and in spite of its close affinity with the cultures of the other countries of Southeast Asia.

Myanmar was one of the first areas in Southeast Asia to receive Buddhism, and by the 11th century it had become the centre of the Theravāda branch of Buddhism. The faith was patronized by the country's leadership, and it became the ideological foundation of the Myanmar state that blossomed at Pagan on the dry central plains.

The origins of civilization in Myanmar. The Irrawaddy River, flowing southward through the entire north-south length of modern-day Myanmar, divides the country in two, and its valley forms the central plain. In addition, the region has long been divided culturally into northern Myanmar (or Upper Burma), the areas north of the Irrawaddy delta, and southern Myanmar (or Lower Burma), the delta and peninsular areas.

The first human settlements in Myanmar appeared some 11,000 years ago in this valley. The stone and fossilized-wood tools used by these people have been named Anyathian, from Anyatha (another term for Upper Burma); little else, however, is known of these people. A discovery in 1969 by workers from the government's Department of Archaeology of some cave paintings and stone tools in the eastern part of present-day Shan state shows that that area, too, had Paleolithic and Proto-Neolithic settlements, the culture of which was similar to the Hoabinhian culture that was widespread in the rest of Southeast Asia. Crude shards and ring stones found at the site appear to have been attached to stonecutting tools to make them more suitable for digging. The woodcutting tools in the find probably were used to clear patches of forest for cultivation, which would indicate that the shift from gathering to agriculture had already begun.

The Pyu state. Between the 1st century BC and the 9th century AD, speakers of Tibeto-Burman languages known as the Pyu were establishing city-kingdoms in Myanmar at Binnaka, Mongamo, Śrī Kṣetra, and Halingyi. There long had been a trade route between China and India that passed through northern Myanmar and then across the Chindwin River valley. In AD 97 and 121 Roman embassies to China chose the overland route through Myanmar for their journey. The Pyu, however, provided an alternative route down the Irrawaddy to Śrī Kṣetra and then by sea westward to India and eastward to insular Southeast Asia, from where the China trade connected with the portage routes and the islands. Chinese historical records noted that the Pyu claimed sovereignty over 18 kingdoms, many of them in the southern portions of Myanmar.

Pyu life

The same Chinese records emphasized the humane nature of Pyu government and the elegance and grace of Pyu life. Fetters, chains, and prisons were unknown, and punishment for criminals was a few strokes with the whip. The men, gaily dressed in blue, wore gold ornaments on their hats, and the women wore jewels in their hair. The Pyu lived in houses built of timber and roofed with tiles of lead and tin; they used golden knives and utensils and were surrounded by art objects of gold, green glass, jade, and crystal. Parts of the city walls, the palace, and the monasteries were built of glazed brick. The Pyu also appeared to have been Buddhists of the Sarvāstivāda school. Their architects may have developed the vaulted temple, which later found its greatest expression at Pagan during its golden age. Pyu sons and daughters were disciplined and educated in monasteries or convents as novices. In the 7th century the Pyu shifted their capital northward to the dry zone, leaving Śrī Ksetra as a "secondary" centre to oversee trade.

The Mon. To the south of the Pyu lived the Mon, who were speakers of an Austro-Asiatic language. The Mon were closely related to the Khmer, who lived to the east of the Mon in what is now Cambodia. The capital of the Mon probably was the port of Thaton, which was

located northwest of the mouth of the Salween River and not far from the portage routes of the Malay Peninsula; through this window to the sea the Mon saw India, in its full glory, under the Gupta empire. Earlier, in the 3rd century BC, the great Mauryan emperor Aśoka apparently had sent a mission of Buddhist monks to a place called Suvarnabhūmi (the Golden Land), which is now thought to have been in the Mon region of the Isthmus of Kra. The ancient monastic settlement of Kelasa, situated near Thaton in southern Myanmar and claimed by Burmese and Mon chronicles to have been founded by Aśoka's missionaries, was mentioned in early Sinhalese records as being represented at a great religious ceremony held in Sri Lanka in the 2nd century BC.

With the expansion of Indian commerce in Southeast Asia between the 1st and 4th centuries AD, Thaton's prosperity and importance increased. Indian merchants and seamen came to Thaton as traders rather than as conquerors or colonists. The number of Indians was never great, and their settlements were of a commercial, not military, nature. As a result, Indian culture was readily accepted by the Mon.

Mon culture

The Mon culture was not abandoned or displaced, and the Mon blended the old with the new. They integrated many of their own beliefs into those of Theravāda Buddhism, which arrived in Southeast Asia already replete with folk beliefs. The power and prestige of the Mon kingship were enhanced by the notions of kingship found in India. The Mon developed a new art of sculpture by blending native traditions with Gupta conventions of iconography. They built stupas (Buddhist ceremonial mounds) according to Indian models, which were adapted, however, to indigenous aesthetic tastes. The Mon subsequently became one of the most culturally advanced peoples in Southeast Asia. They assumed the role of teachers to their neighbours, spreading Theravāda Buddhism and their new culture over the entire region.

The Mon centre eventually shifted to Pegu, located on the Pegu River, about 50 miles northeast of present-day Yangôn (Rangoon). From there the Mon were able to control the trade of southern Myanmar.

The Pagan kingdom (849–1287). *The advent of the Pagan Burmans.* Another group of Tibeto-Burman speakers, the Burmans, also had become established in the northern dry zone. They were centred on the small settlement of Pagan on the Irrawaddy River. By the mid-9th century, Pagan had emerged as the capital of a powerful kingdom that would unify Myanmar and would inaugurate the Burman domination of the country that has continued to the present day.

During the 8th and 9th centuries the kingdom of Nanchao became the dominant power in southwestern China; it was populated by speakers of Lolo (or Yi), a Tibeto-Burman language. Nanchao mounted a series of raids on the cities of mainland Southeast Asia in the early decades of the 9th century and even captured Hanoi in 861. The Mon and Khmer cities held firm, but the Pyu capital of Halingyi fell. The Burmans moved into this political vacuum, establishing Pagan as their capital city in 849.

By that time the Mon apparently had become supreme in southern Myanmar. They may have occupied the whole of the region and controlled the port of Bassein (now Pathein) in the west and the city of Pegu in the centre. They could have stepped into the vacuum caused by the destruction of the Pyu kingdom, but their power was linked to the trade of southern Myanmar and not with the agrarian-based economy of northern Myanmar.

The unification of Myanmar. Nanchao acted as a buffer against Chinese power to the north and allowed the infant Burman kingdom to grow. The Burmans learned much from the Pyu, but they were still cut off from the trade revenues of southern Myanmar. Theravāda Buddhism had disappeared from India, and in its place were Mahāyāna Buddhism and a resurgent Hinduism.

Reign of Anawrahta

In 1044 Anawrahta came to the throne at Pagan and began the unification process in Myanmar that would recur in cyclic fashion until the British conquered the country in 1886. Anawrahta first strengthened his defenses on the north—the "front door" of Myanmar—and created alliances through marriage with the neighbouring Shan to the east. He then harnessed the economic resources of northern Myanmar by repairing old irrigation works and building new ones. Finally, he declared himself the champion of Theravāda Buddhism and used that ideology to justify his conquest of southern Myanmar, which was accomplished with the defeat of the Mon city of Thaton in 1057.

Thus, by the mid-11th century the core of modern-day Myanmar had been united into a single kingdom centred at Pagan, and Myanmar's longest-surviving dynasty had been established. Anawrahta's work was continued by his great commander Kyanzittha (ruled 1084–*c.* 1112) and by another great ruler, Alaungsithu (ruled *c.* 1112–*c.* 1167). Pagan's consolidation of the Irrawaddy valley southward to the ports of southern Myanmar divided most of mainland Southeast Asia into two great empires, Khmer and Burman. Anawrahta's dynasty of kings lasted until the 13th century. By that time, their great temples had been built, and their message of Theravāda Buddhism had been carried not only to the Shan but also to the Khmer.

Centuries of temple building and of donations of land and manpower to the tax-exempt sangha (monkhood), however, had diverted much of the state's most valuable resources. Yet, the legitimacy of state and society depended on continued patronage of the sangha. As a result, Pagan had been weakened by the end of the 13th century, precisely when the Mongols threatened. Pagan had lost its northern buffer in the early 1250s when Nanchao was destroyed and subjugated by the Mongols under Kublai Khan. The Mongols demanded submission by and tribute from Pagan, which refused to comply. It is not clear if the Mongol armies actually reached Pagan, but by 1300 Pagan no longer was the centre of power in Myanmar.

Pagan state and society. Pagan was a fabulous kingdom even to its contemporaries; although he did not visit it, the Venetian traveler Marco Polo was impressed by the tales of its splendour that were recounted to him. By the time of its conquest, Pagan had an estimated 3,000 to 4,000 temples and monasteries. Hundreds of these still stand today and testify to the prosperity of its people and the richness of its culture. The conquest of the Mon kingdom of Thaton was the foundation of both Pagan's economy and its culture, for it delivered into Burman hands all the ports of the country and the core of artisans who built Pagan's magnificent temples. These artisans were paid in wages of gold and silver, as well as in kind (food and horses and elephants). Their clothing, shelter, health, comfort, and safety were the responsibility of their employers (as evidenced by contemporary inscriptions that provide the details connected with these topics).

Cultural achievements

The Mon craftsmen, artisans, artists, architects, goldsmiths, and wood-carvers who were captured at Thaton and taken to Pagan taught their skills and arts to the Burmans. Mon monks and scholars taught the Burmans the Pāli language and the Buddhist scriptures, and the Burmans soon became scholars themselves, making Pagan the centre of Theravāda learning. Some of their religious commentaries came to be accepted as part of the Pāli canon by other Theravāda countries. The women of Pagan took part in all these activities, particularly in the building and endowment of temples and monasteries.

The people of Pagan largely were devout and orthodox, and they made Buddhism their way of life while still retaining animistic and other unorthodox beliefs. They established the principles underlying religion, government, and society that later generations accepted almost without change. Thus, it is to Pagan that the designation the "classical age" of Myanmar is given.

Myanmar from the end of Pagan to 1885. *The first Ava dynasty.* After the decline of Pagan as a political centre, three small centres of power emerged by the first decade of the 14th century from polities that once had been under Pagan suzerainty. The political situation remained fragmented until 1364, when Ava became the seat of authority. It was located in the northern Irrawaddy valley at the entrance to the rice-producing region of Kyaukse, near present-day Mandalay. The kings of Ava resurrected the traditions of Pagan, encouraging scholarship and learning

and making the period a great age of Burmese literature. Without a northern buffer, however, they could not control the coasts for any length of time and were thus deprived of shipping revenues.

After the decline of Pagan's political authority, the Mon reestablished themselves at Pegu, and by the 15th century they were experiencing their own golden age under rulers like Dhammazedi (reigned 1472–92). Pegu became a major centre of Theravāda scholarship and of commerce in Southeast Asia, attributes that protected it from conquest. In 1527 Ava was sacked by the Shan, who had been moving southward down the Irrawaddy and Chao Phraya valleys since Nanchao had been destroyed. The Ava refugee population fled south to Toungoo, a city on the Sittang River that had been a seat of Pagan and Ava authority.

The Toungoo dynasty. In 1531 a dynasty was established at Toungoo by Tabinshwehti, and by then the new kingdom had become powerful enough to conquer northern Myanmar from the Shan and southern Myanmar from the Mon. In 1511 the great trading entrepôt of Malacca (Melaka) on the Malay Peninsula had been conquered by the Portuguese, which led to a renewal of interest in trade in Myanmar's coastal waters. Tabinshwehti transferred his capital to Pegu in order to tap this commercial potential, and he attempted to unite Burmans, Mon, and Shan into a single nation. He died in 1550, however, and was succeeded by his brother-in-law, Bayinnaung.

Meanwhile, the Shan in the Chao Phraya valley had consolidated their power under the Tai kingdom of Ayutthaya (Ayudhia, or Siam), and they also recognized the potential value of controlling this renewed commercial activity. In addition, the Ming Chinese were active in Southeast Asian waters during the 14th and 15th centuries, further contributing to the growth of economic activity there. Thus, Pegu and Ayutthaya became rivals.

The rule of Bayinnaung

Bayinnaung twice marched on Ayutthaya and conquered the entire Chao Phraya valley by 1569, using Portuguese mercenaries and Portuguese cannon to accomplish his goals. Bayinnaung's wars exhausted Myanmar's resources, however, and after his death the kingdom began to break up. Manipur, a Hindu princely state to the northwest of Myanmar that had been subjugated in 1560, declared itself independent, and Ayutthaya also regained its independence. Toungoo, joined by Arakan (Rakhine), proceeded to ravage Pegu. The Portuguese founded a small centre of power at Syriam, which was located on the Pegu River across from the site of present-day Yangôn. At this point, the rulers of Toungoo decided to return to the predictability, security, and comfort of the agrarian dry zone of northern Myanmar.

By the end of the 16th century, Ava had been resurrected and the second Ava dynasty established. Bayinnaung's grandson, Anaukpetlun, reunited Myanmar once more by 1613. His successor, Thalun, reestablished the principles of the classical Myanmar state created half a millennium earlier at Pagan. Heavy religious expenditures, however, weakened Ava politically, much as they had done in Pagan. In the meantime, southern Myanmar had been rejuvenated by the new commercial activity spurred by the British and Dutch. Pegu had grown stronger while Ava had been preoccupied with resurrecting the "heartland." Finally, Pegu rose in rebellion, encouraged by the French in India. Assailed internally as well as externally, Ava fell in 1752.

The Konbaung dynasty. It was soon apparent, however, that only the centre of power had been destroyed, not the system nor the wherewithal for power; before the year had ended, a popular Burman leader, Alaungpaya, drove Pegu's forces out of northern Myanmar and regained the Shan states. By 1759 he had regained Manipur and defeated Pegu. The Siamese became alarmed and attempted to rouse the Shan chiefs to rebel. Alaungpaya retook Tenasserim, the site of the old portage kingdoms, and invaded Siamese territory. Although Alaungpaya's invasion failed and he himself died during the retreat in 1760, the Myanmar now felt that, unless the Siamese were conquered, the coastal cities of southern Myanmar could not be retained. Alaungpaya's son, Hsinbyushin, sent his armies into Siam in 1766, and they captured Ayutthaya in 1767. China, alarmed over the growing power of Myanmar, invaded the country four times during the period 1766–69, without success.

Myanmar then conquered Arakan and occupied the princely state of Assam to the northwest of Manipur, thus coming face to face with British power in India. The result was the First Anglo-Burmese War (1824–26), in which the Siamese fought on the British side. Myanmar eventually had to sue for peace and lost Assam, Manipur, Arakan, and Tenasserim.

The Second Anglo-Burmese War (1852) was provoked by the British, who wanted access to the teak forests in and around Pegu and also wanted to secure the gap in their coastline stretching from Calcutta to Singapore; it resulted in the British annexation of Pegu province, which they renamed Lower Burma. As the British became increasingly interested in the legendary trade with China through its back door—as well as in the teak, oil, and rubies of northern Myanmar—they waited for a suitable pretext to attack. In 1885 Britain declared war on Myanmar for the third and final time. To meet the criticism of their action that arose in Parliament, the British government gave the excuses that the last independent king of Myanmar, Thibaw (ruled 1878–85), was a tyrant and that he was conspiring to give France greater influence over the country. Neither of these charges seems to have had much foundation.

The Third Anglo-Burmese War

The administration of traditional Myanmar. The king was the chief executive and the final court of appeal, but there were checks on his power. He could not make laws, only issue administrative edicts that might or might not be upheld after his death. Custom was a strong and recognized source for proper behaviour, along with codified bodies of civil and criminal law called, respectively, the Dammathat and Rajathat.

The king, as the patron of Buddhism and the head of state, was expected to be both a conqueror and one who renounced the world. Buddhist monks were formally organized into a church that was headed by a primate who, although appointed by the king, sometimes proved to be the king's sternest critic. Although monks technically were supposed to remain outside the sphere of politics, they gave sanctuary to political exiles. Monasteries also served as schools for all children, and monks educated the people and molded public opinion regarding the state and the king. Because the state and church owned virtually all the productive land in Myanmar, there were no landed, hereditary nobles who could weaken the power of the state. The king's officials were appointed, and their appointments could lapse with the king's death.

The Hluttaw, or Hlutdaw ("Place of Release"), was the centre of government. It had several integrated functions—including fiscal, executive, and judicial responsibilities—and it was the final court of appeal; in theory and often in practice the king presided over its deliberations. All proclamations and appointments that were made by the king became valid only when orders giving effect to them were issued by the Hluttaw.

Every province had a governor, to whom were delegated certain powers by the Hluttaw. There always was a right of appeal against all decisions of the governer to the Hluttaw. Local government was in the hands of hereditary "headmen," who were advised by village elders. The position of the headman was officially confirmed by the king.

The British in Burma (1885–1948). The Third Anglo-Burmese War lasted less than two weeks during November 1885. The Myanmar people never expected the speed with which the capital would be taken. The hopelessly outmatched royal troops surrendered quickly, although armed resistance continued for several years. The Myanmar also believed that the British aim was merely to replace King Thibaw with a prince who had been sheltered and groomed in India for the throne. This belief seemed to be confirmed when the British commander called upon the High Court of Justice to continue to function. The British finally decided, however, not only to annex all of northern Myanmar (which they called Upper Burma) as a colony but also to make the whole country a province of

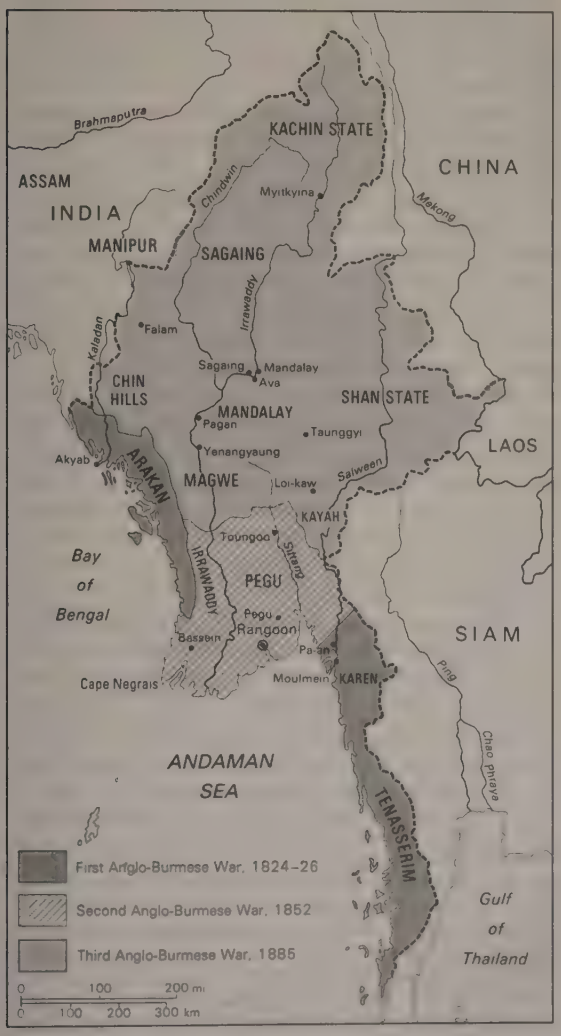
Burma as a province

India (effective Jan. 1, 1886). Rangoon became the capital of the province, after having been the capital of British Lower Burma.

The initial impact of colonialism. This chain of events was a bitter blow to Myanmar society. The loss of independence was painful enough; worse still were the British decisions to eliminate the monarchy—in the process sending Thibaw into exile—and to stay out of religious affairs, thus depriving the church of its traditional status and official patronage. The demise of these twin pillars of Myanmar society was perhaps the most devastating aspect of the colonial period.

Many refused to accept the British victory as final and resorted to guerrilla warfare against the British army of occupation. The guerrillas were led mainly by former officers of the disbanded royal army, former officials (including village headmen), and royal princes, and they considered themselves to be royal soldiers still fighting the Third Anglo-Burmese War. To the British, however, the war had ended legally with the annexation of the kingdom; those opposing them, therefore, were considered rebels and bandits. For the next five years the British military officers acted as both judge and jury in dealing with captured guerrillas. Villagers who aided the rebels also were sternly punished. British troops carried out mass executions and committed other atrocities.

As the guerrillas fought on, the British adopted a "strategic-hamlet" strategy: villages were burned, and families who had supplied villages with their headmen were uprooted from their homes and sent away to Lower Burma (which had been under British control since the Second Anglo-Burmese War). Strangers loyal to the British were appointed as headmen for the new villages established by the British. The guerrillas resorted to desperate measures against the new village officials. By 1890, with more than 30,000 British troops engaged in the campaign, the military part of the struggle was over.



British territorial acquisitions in Burma.

The religious dilemma. The colonial period was one of relative civil order, but it also was one of great social disintegration. Chief among the reasons for this was the British-imposed separation of church and state. The British did not wish to touch the issue of religion—given their experience in India that had led to the revolt of 1857–59—and thus were unwilling to patronize Burmese Buddhism as the monarchy had done.

Under the monarchy, church and state had shared a symbiotic relationship. Royal patronage of Burmese Buddhism had included both financial and moral support, and this had extended legitimacy and authority to the church. The king had had the right to appoint the primate, who exercised supervision and discipline among the ranks of the clergy. In addition, the king had been given the right to attach two royal officials to the primate: a commissioner of ecclesiastical lands and an ecclesiastical censor. The duty of the commissioner had been to see that ecclesiastical lands were exempted from payment of taxes, at the same time ensuring that bogus and illegal endowments did not escape taxation. The duty of the censor had been to maintain a register of monks, which had given the king indirect control over the clergy. The power to defrock a wayward monk had rested largely with the primate, but the same result could be achieved if the king declared the monk to be "impure," which was one of the king's prerogatives. This arrangement was designed to prevent the abuse of the exemptions granted to the clergy.

The British refusal to heed a plea by the clergy and church elders to continue the traditional church-state relationship resulted in the decline of the sangha and its ability to instill discipline in the clergy. This, in turn, lowered the prestige of the clergy and contributed to the rise of secular education and of a new class of teachers, depriving the sangha of one of its primary roles. Added to this, the colonial government of India founded secular schools teaching in both English and Burmese and encouraged foreign Christian missions to found schools by offering them financial assistance. Many mission schools were founded; parents were compelled to send their children to these schools, as there were no realistic alternatives. The teachers were missionaries, and the lessons they gave were marked by repeated criticism of Buddhism and its culture. In the government schools the first teachers, British and Indian, were mere civil servants, unable and unwilling to continue the older traditions.

Education under the British

The colonial economy. The traditional Myanmar economy had been one of redistribution—an early form of the modern command economy—a concept that was embedded in society, religion, and politics. Prices of the most important commodities were set by the state, and in general the mechanism of supply and demand was relatively unimportant. Agrarian self-sufficiency was vital, while trade was only of secondary importance. The British impact on this system proved disastrous, as Burma's economy became part of a vast colonial export economy tied to global market forces. The cause of this was not so much the usual economic exploitation of a country by a foreign ruler as it was the effects of an economy designed to benefit the colonial power.

The British dream of a golden road to China through Burma could not be realized, but the opening of the Suez Canal in 1869 created a much higher international demand for Burma's rice than had previously existed. The Irrawaddy delta was swiftly cleared of its mangrove forests and in a matter of decades became covered with rice fields. Even in 1857 the price of rice had increased 25 percent; by 1890 the price had more than doubled from 1857 rates, and it continued to increase until the worldwide economic depression of the 1930s. The area of productive rice fields in Lower Burma rose from approximately 60,000 acres (24,000 hectares) in the 1850s to nearly 10,000,000 acres in 1939. This tremendous increase in production created a significant shift in population from the northern heartland to the delta, shifting as well the basis of wealth and power.

In order to prepare the land for cultivation, however, the farmers had to borrow capital from Indian moneylenders from Madras at exorbitant interest rates. The British banks would not grant mortgage loans on rice land, and the British government had no policy for establishing land-mortgage banks or for making agricultural loans. Prevailing prices were high in the international market, but the local price was kept down by a handful of British firms that controlled wholesale trade and by Indian and Chinese merchants who controlled retail trade. With land values and rice prices soaring, the Indian moneylenders foreclosed mortgages at the earliest opportunity.

Plight of the farmers

The dispossessed farmers could not find employment even on their lost lands because, with a higher standard of living, they could not compete with the thousands of Indian labourers who came to Burma. Burmese villagers, unemployed and lost in a disintegrating society, took to petty theft and robbery and soon acquired the reputation of being lazy and undisciplined. The level of dysfunction in Burmese society was revealed by the dramatic rise in homicides: during the early decades of the 20th century, Burma's annual homicide rate was second only to that of the United States.

Thus, although the Burmese economy and transportation infrastructure developed rapidly from 1890 to 1900, the majority of Burmese people did not benefit from it. A railway had been built through the entire valley of the Irrawaddy, and hundreds of steamboats plied the length of the river; but the railway and the boats belonged to British companies. Roads had been built by the government, but they were meant for the swift transport of troops. A British company worked the ruby mines until they became nearly exhausted. The extraction of oil and timber was monopolized by two British firms. The balance of trade was always in favour of Burma, but that meant little to Burmese people or society.

The emergence of nationalism. Those Burmese who attended the new schools managed to gain admission to the clerical grades of government service, but even in those lower grades they encountered competition from Indians. Because science courses were not available, the professions of engineering and medicine were closed to the Burmese. Those who advanced to the government liberal arts college at Rangoon entered the middle grades of the civil service, while a few went on to London to study law. When these young barristers returned to Burma, they were looked upon by the people as their new leaders. Their sojourn in the liberal atmosphere of London had convinced these new leaders that some measure of political independence could be regained by negotiation.

The new leaders first gave their attention to the national religion, culture, and education. In 1906 they founded the Young Men's Buddhist Association (YMBA) and through it began establishing a number of schools supported by private donations and government grants-in-aid (the YMBA was not antigovernment). Three years later the British, attempting to pacify the Indian National Congress, introduced some constitutional reforms in India. Only a few minor changes were made in the Burmese constitution, but these confirmed the young leaders' faith in British liberalism. In 1920, however, when it was found that Burma would be excluded from new reforms introduced in India, the barristers led the people in a nationwide protest, which involved a boycott of British goods.

Also in 1920 Rangoon College was raised to the status of a full university; yet, because its control was vested in a body of government nominees, its students went on strike. Younger schoolboys and schoolgirls followed suit. The strikers camped in the courtyards of monasteries, reviving memories of days when education was the concern of the monks. The general public and the Buddhist clergy gave full support to the strike. The University Act was amended and the strike settled, but many strikers refused to go back to mission and government schools. The YMBA schools, now calling themselves "national" schools, opened their doors to the strikers.

Constitutional reforms were finally granted in 1923, but the delay had split the leaders, some of whom, like the masses, were beginning to doubt whether political freedom could be attained by peaceful protest. In the University of Rangoon itself, students began to resent their British professors. A radical student group began organizing protests, which came to be known as the Thakin movement. The name for this movement was purposely ironic: the Burmese word *thakin* ("master") was the term that the Burmese were required to use when addressing the British. (M.H.Au./M.A.A.-T.)

The Thakin movement

In late 1930 Burmese peasants, under the leadership of Saya San, rose in rebellion. Armed only with swords and sticks, they resisted British and Indian troops for two years. The young Thakins, though not involved in the rebellion, won the trust of the villagers and emerged as leaders in place of the British-educated Burmese elite. In 1936 university students again went on strike, and two of their leaders, Thakin Nu (later called U Nu) and Aung San, joined the Thakin movement. In 1937 the British government separated Burma from India and granted it a constitution, but the masses interpreted this as proof that the British planned to exclude Burma from the next phase of Indian reform.

World War II and after. When World War II broke out in Europe in 1939, the Burmese leaders wanted to bargain with the government before giving their support to the British. A warrant was issued for the arrest of Aung San, but he escaped to China, where he attempted to contact radical groups for support. Japanese assistance was offered instead. Aung San returned to Burma in secret, recruited 29 young men, and took them to Japan, where the "Thirty Comrades" received military training. The Japanese promised independence for Burma; hence, when Japanese troops reached Bangkok (Thailand) in December 1941, Aung San announced the formation of the Burma Independence Army (BIA). The Japanese advanced into Burma and by the end of 1942 had occupied the country. They subsequently disbanded the BIA and formed a smaller Burma Defense Army, with Aung San still as commander.

Ba Maw, the first prime minister under the 1937 constitution and later the leader of the opposition, was appointed head of state by the Japanese, with a cabinet including Aung San and Thakin Nu. In 1943, when the tide of battle started to turn against them, the Japanese declared Burma a fully sovereign state. The Burmese government, however, was still a mere facade, with the Japanese army ruling. Meanwhile, Aung San had contacted Lord Mountbatten, the Allied commander in Southeast Asia, as early as October 1943 to offer his cooperation, and in March 1945 Aung San and his army—renamed the Burma National Army (BNA)—joined the British side.

During the war Aung San and the Thakins formed a coalition of political parties called the Anti-Fascist Organization—after the war renamed the Anti-Fascist People's Freedom League (AFPFL)—which had wide popular support. After the defeat of the Japanese in Burma in May 1945, the British military administration and members of the prewar government who had returned from exile demanded that Aung San be tried as a traitor. Mountbatten, however, recognized the extent of Aung San's hold on the BNA and on the general populace, and he hastily sent the more conciliatory Sir Hubert Rance to head the administration. Rance regained for the British the trust of Aung San and the general public. When the war ended, the military administration was withdrawn, and Rance was replaced by the former civilian governor, who formed a cabinet consisting of older and more conservative politicians. The new administration arrested Aung San and charged him with treason. Surprised and angered, the Burmese people prepared for rebellion, but the British government in London wisely replaced the governor with Rance. Rance formed a new cabinet, including Aung San, and discussions for a peaceful transfer of power began. These were concluded in London in January 1947, when the British agreed to Burma's independence; by June the decision also had been made to leave the Commonwealth.

The AFPFL

The communist and conservative wings of the AFPFL were dissatisfied with the agreement. The communists broke away and went underground, and the conservatives went into opposition. In July Aung San and most mem-

bers of his cabinet were assassinated by gunmen sent by U Saw, a former prime minister and now a conservative. Rance asked Thakin Nu to form a new cabinet. A new constitution was written, and on Jan. 4, 1948, Burma became a sovereign, independent republic.

The country since independence. *Parliamentary government.* With its economy shattered and its towns and villages destroyed during the war, Burma needed peace. A foreign policy of neutrality was decided upon, but, because of internal strife, no peace resulted. The communists were the first insurgents, followed by some of Aung San's veterans and then the Karen, the only ethnic minority on the plains; but the other minorities—Chin, Kachin, and Shan—who had been ruled separately by the British but who had enthusiastically joined the union, stood firm.

A division of Chinese Nationalist troops occupied parts of the Shan Plateau after their defeat by the Chinese Communists in 1949; and, because of the general support given to Nationalist China (Taiwan) by the United States, Myanmar stopped accepting U.S. aid and rejected all foreign aid. At the United Nations Myanmar endeavoured to show impartiality. It was one of the first countries to recognize Israel and also the People's Republic of China.

By 1958 Burma was well on the road to internal peace and economic recovery, but the ruling AFPFL had become divided by personal quarrels between U Nu and his closest associates. Amid rumours of a military takeover, U Nu invited the army chief of staff, Ne Win—who had been a Thakin, one of the Thirty Comrades, and Aung San's second in command—to take the premiership. Ne Win established internal security, stabilized the military situation, and prepared the country for general elections, which took place in February 1960. U Nu was returned to office with an absolute majority.

The socialist takeover. In March 1962 Ne Win led a coup d'état and arrested U Nu, the chief justice, and several cabinet ministers. His justification for this was that it was a means of keeping the union from disintegrating. He suspended the 1947 constitution and ruled the country with a Revolutionary Council consisting of senior military officers. Ne Win's stated purpose was to make Burma a truly socialist state. A one-party (Burma Socialist Programme Party [BSPP]) system was established, and measures were introduced to decentralize the administration. In April 1972 Ne Win and other members of the Revolutionary Council retired from the army, but they retained their positions of power in the BSPP. Land had been nationalized in U Nu's administration, and now much of the country's commerce and industry were nationalized as well. These measures did not improve the economy, however, particularly as investment in agriculture generally was sacrificed in favour of industrial growth.

The constitution of 1974

U Ne Win (as he was called after leaving the army) had promised a new constitution, and in September 1971 representatives of the party's central committee, of the country's various ethnic groups, and of other interest groups were appointed to draft a document. A referendum to ratify the new constitution was held in December 1973, with more than 90 percent of eligible voters signifying approval, and the constitution was promulgated in January 1974. Elections to the People's Assembly (Pyithu Hluttaw)—the supreme legislative, executive, and judicial authority—and to local People's Councils were held early in 1974; the new government took office in March with U Ne Win as president.

After the establishment of the new political organization, Burma's economy grew steadily at a moderate pace. A notable policy change was a partial relaxation of the ban on foreign financial aid, and considerable funding was received from the Asian Development Bank and the International Bank for Reconstruction and Development. By the early 1980s, however, growth increasingly was being hindered by mounting trade deficits caused largely by falling commodity prices and rising external debt payments. A series of economic reforms were proposed in 1987–88 that would reverse the socialist policies enacted in the early 1960s. Chief among these were the further encouragement of foreign investment and a considerable liberalization of foreign trade.

Communist and ethnic insurgency continued in the eastern and northern parts of the country throughout the BSPP period. In May 1980 U Ne Win offered full amnesty to all political insurgents inside or outside Burma who reported to authorities within a 90-day period. Most notable among those accepting was U Nu, who, after having gone into exile in India in 1969, returned to enter a Buddhist monastary. Most insurgents, however, chose to continue opposing the government, and repeated attempts by government troops to suppress them met with only limited success. After four decades, insurgency had become a way of life.

The military regime. U Ne Win retired as president and chairman of the Council of State in November 1981 but remained in power until July 1988, when he resigned as chairman of the BSPP amid violent protests. Student and worker unrest had erupted periodically throughout the 1980s, but the intensity of the protests in the summer of 1988 made it seem as if the country was on the verge of revolution. In September the armed forces, led by General Saw Maung, seized control of the government. The military moved to suppress the demonstrations, and thousands of unarmed protesters were killed. Martial law was imposed over most of the country, and constitutional government was replaced by a new body called the State Law and Order Restoration Council (SLORC). Saw Maung became chairman of the SLORC and prime minister.

The SLORC implemented the economic reforms drafted by the previous government and called for elections for a new Constituent Assembly to revise the 1974 constitution.

Election of 1990

In May 1990 Myanmar held its first multiparty elections in 30 years. Of the dozens of parties that participated, the two most important were the government's National Unity Party (NUP), which was the successor of the BSPP, and an opposition coalition called the National League for Democracy (NLD). The result was a landslide victory for the opposition NLD.

The SLORC, however, would not permit the Constituent Assembly to convene immediately. In addition, the NLD's leaders, U Tin U, a former colleague of U Ne Win, and Daw Aung San Suu Kyi, the daughter of the nationalist leader Aung San, had been under house arrest since July 1989, while another leader, U Sein Win, remained in exile in the West. International condemnation of the military regime was strong and widespread, both for its bloody repression of the demonstrations in 1988 and for its actions in connection with the 1990 elections. Worldwide attention remained focused on Myanmar after Aung San Suu Kyi was awarded the Nobel Peace Prize in 1991. In April 1992 Saw Maung was replaced as chairman of the SLORC and prime minister by General Than Shwe. The SLORC permitted the new assembly to convene for the first time in January 1993. (M.A.A.-T.)

For later developments in the history of Myanmar, see the BRITANNICA BOOK OF THE YEAR.

Singapore

The Republic of Singapore (Chinese [Wade-Giles transliteration]: Hsin-chia-p'o Kung-ho-kuo; Malay: Republik Singapura; Tamil: Singapore Kudiyarasu) is a city-state situated at the southern tip of the Malay Peninsula, about 85 miles (137 kilometres) north of the Equator. It consists of the diamond-shaped Singapore Island and some 60 small islets, for a combined area of about 240 square miles (622 square kilometres); the main island occupies all but about 18 square miles of this territory. The main island is separated from Peninsular Malaysia to the north by Johor Strait, a narrow channel crossed by a road and rail causeway that is more than half a mile long. The southern limits of the state run through Singapore Strait, where outliers of the Riau-Lingga Archipelago—which forms a part of Indonesia—extend to within 10 miles of the main island.

Singapore is the largest port in Southeast Asia and one of the busiest in the world. It owes its growth and prosperity to its focal position at the southern extremity of the Malay Peninsula, where it dominates the Strait of Malacca, which connects the Indian Ocean to the South China Sea. Once a British colony and now a member of

the Commonwealth, Singapore first joined the Federation of Malaysia on its formation in 1963 and subsequently seceded to become an independent state on Aug. 9, 1965.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Nearly two-thirds of the main island is less than 50 feet (15 metres) above sea level. Timah Hill, the highest summit, has an elevation of only 531 feet (162 metres); with other peaks, such as Panjang and Mandai hills, it forms a block of rugged terrain in the centre of the island. To the west and south are lower scarps with marked northwest-southeast trends, such as Mount Faber. The eastern part of the island is a low plateau cut by erosion into an intricate pattern of hills and valleys.

Geology

These physical units reflect their geologic foundations: the central hills are formed from granite rocks, the scarp lands from highly folded and faulted sedimentary rocks, and the eastern plateau from uncompacted sands and gravels.

Drainage and soils. A dense network of short streams drains the island, but floods are locally severe because the streams have low gradients and because of excessive water runoff from cleared land. Many streams, especially those draining northward, have broad mangrove-fringed estuaries that extend far inland. None of the soils is even reasonably fertile, but those derived from the granites tend to be better than most. Soils developed from the sedimentary rocks are variable, but many contain hardpans (compacted layers) that restrict plant roots and impede soil drainage. The soils of eastern Singapore are extremely infertile. All have suffered extensive degradation through erosion as a result of generations of careless human exploitation.

Climate. Singapore is in the equatorial monsoon region of Southeast Asia, and its climate is characterized by uniformly high temperatures and nearly constant precipitation throughout the year. The average monthly temperature varies from about 81° F (27° C) in June to 77° F (25° C) in January. The daily range is somewhat greater, averaging about 13° F (7° C). Singapore's maritime location and constant humidity, however, keep maximum temperatures relatively moderate: the highest temperature ever recorded was only 97° F (36° C).

The seasons are defined by the relative incidence of rainfall, which, in turn, is determined by the movements of the monsoon air masses. The wettest and windiest period is during the northeast monsoon (November–March), with rainfall reaching an average monthly high of more than 10 inches (250 millimetres) in December. Conversely, the period of the least amount of rainfall and the lightest winds is during the southwest monsoon (May–September), with rainfall dropping to a monthly low of less than 7 inches in July. April and October are intermonsoonal periods characterized by sluggish air movements and intense afternoon showers and thunderstorms. Altogether, Singapore's precipitation averages about 95 inches annually, and rain falls somewhere on the island every day of the year.

Plant and animal life. Little remains of the original vegetation or animal life, except for a few thousand acres of evergreen rain forest preserved around catchment areas. Some mangrove vegetation survives in the Kranji area on the northwest side of the island, but elsewhere tracts of scrub or cogon grass (called *lalang* locally) are common. Many exotic plants have been introduced for ornamental use. The largest native animals are the long-tailed macaque (an Asian species of monkey), the slow loris (a large-eyed tailless nocturnal lemur), and the scaly anteater. Birds are numerous, especially those like the Indian mynah bird, the brahminy kite (a kite with reddish brown plumage and a white head and breast), and the house swallow that have adapted to a symbiotic relationship with humans. Reptiles, such as cobras and lizards, also are common. Fringing coral reefs with their associated fish and wildlife occur around many parts of the coast.

Birds

Settlement patterns. The city of Singapore is situated in the southern portion of the main island. Over time, urbanization has blurred the differences between city and country. Built-up areas now cover a large part of the city-state. The older parts of the city have been substantially refurbished, especially along the Singapore River but elsewhere as well. The once-common Chinese shop-house, consisting of living quarters above a commercial establishment, gradually has been disappearing from the city. Instead, the government's Housing and Development Board (HDB) has relocated commerce into separate districts and has created integrated residential communities inhabited by people with a mixture of incomes. About four-fifths of Singapore's population now resides in high-rise HDB flats located in housing estates and new towns. The new towns—such as Woodlands, Tampines, and Yishun—are scattered across the island and are characterized by easy access to places of employment and shopping districts. The traditional Malay kampong settlements—consisting of stilt houses built along the shoreline—are declining in number and are now found only in select rural areas.

The people. *Ethnolinguistic composition.* The population of Singapore is diverse, the result of considerable past immigration. Chinese predominate, making up more than three-fourths of the total. Malays are the next largest ethnic group, and Indians the third. None of these three major communities is homogeneous. Among the Chinese, more than two-fifths originate from Fukien province and speak the Amoy dialect, about one-fourth are Teochew from the city of Swatow in Kwangtung province, and a smaller number are from other parts of Kwangtung. The Chinese community as a whole, therefore, speaks mutually incomprehensible dialects. Linguistic differences are less pronounced among the Malays, but the group includes Indonesians speaking Javanese, Boyanese, and other dialects. The Indian group is most diverse, consisting of Tamils (more than half), Malayalis, and Sikhs; it also includes Pakistani and Sinhalese communities.

Official languages

Because of this ethnic diversity, no fewer than four official languages are recognized—English, Mandarin Chinese, Malay, and Tamil. English remains the main medium for administration, commerce, and industry, and it is the primary language of instruction in schools. Mandarin, the official language of China, transcends dialect barriers, and its use is strongly promoted; one-third of the school population is taught in that language. Malay, like English, is widely used for communication among ethnic groups and plays a particularly useful role in view of the close ties between Singapore and Malaysia.

Religions. Religious affiliations reflect ethnic patterns. About two-thirds of all Chinese profess some degree of attachment to Confucianism, Buddhism, or Taoism or to some combination thereof. Virtually all Malays, and some Indians, adhere to Islām, which is the formal religion of about one-sixth of the population. The Christian community has grown rapidly and now constitutes more than 10 percent of the population; nearly all Christians are Chinese. Almost all of the remaining population is Hindu.

Demographic trends. Heavily urbanized, Singapore has a high population density, but it also has been a regional leader in population control. Its birth and population growth rates are the lowest in Southeast Asia. Singapore's high average life expectancy and its low infant-mortality rate reflect high standards of hygiene and access to a superb health care system. The low birth rate and greater longevity of the population have raised the median age, a trend also occurring in other developed nations.

The economy. Singapore, one of the great trading entrepôts of the British empire, has experienced remarkable economic growth and diversification since 1960. In addition to enhancing its position as a world trade centre, it has developed powerful financial and industrial sectors. Singapore has the most advanced economy in Southeast Asia and is often mentioned along with the other rapidly industrializing countries of Asia: Hong Kong, South Korea, and Taiwan. Singapore's economy always has differed from those of the other Southeast Asian countries in that it never has been primarily dependent on the production and export of commodities.

Economic development has been closely supervised by the Singaporean government, and it has been highly dependent on investment capital from foreign multinational corporations. The government holds about three-fourths of all land and is the chief supplier of surplus capital, which is derived largely from contributions to the Central Provident Fund (CPF) social-security savings program. In

addition, the government has attempted to enhance the value and productivity of labour in order to attract investment and boost export competitiveness. This has been accompanied by a strong commitment to education and health. Labour shortages and rising wages have heightened the push for restructuring the economy even more toward higher value-added production.

The rationale for extensive government intervention in economic development has weakened. Official policy relies on market forces, privatization of government enterprise, and more support for domestic private businesses. Union membership has declined as centralized union structures have been replaced by smaller industry- and enterprise-based unions. Greater reliance has been placed on local labour-management negotiations.

Unions

Resources, agriculture, and fisheries. Singapore has few natural resources. There are no natural forests remaining on the island. Only a tiny fraction of the land area is classified as agricultural, and production contributes a negligible amount to the overall economy. Cultivation is intensive, with vegetables and fruits grown and poultry raised for local consumption. The local fishing industry supplies only a portion of the total fresh fish requirement; most of the catch comes from offshore fishing vessels. There also is a small aquaculture industry that raises groupers, sea bass, and prawns. Singapore is a major exporter of both orchids and aquarium fish.

Industry. Since the late 1960s Singapore has pursued a general policy of export-oriented industrialization. In order to attract foreign investment, the economy was liberalized, and a series of incentives were provided to multinational corporations; chief among these was the establishment of free trade zones. Gradually, production has been diversifying from such labour-intensive industries as textiles to high-technology activities like the manufacture of electronics and precision equipment and oil refining, which yield a much higher added value to production.

Services and tourism. Singapore has been able to emphasize its comparative advantage in knowledge-intensive activities—especially communications and information and financial services—which are less dependent on foreign investment. Higher productivity and research and development are encouraged through schemes that provide investment credits and allowances. An effective economic strategy has been to invest local funds abroad and simultaneously to export management skills. Singapore has sought to recruit skilled people, particularly Chinese from Hong Kong, the United States, and China.

Tourism is important to Singapore's economy. The city's central location in Southeast Asia and its excellent air-transport facilities have been augmented by massive investments in hotels and shopping centres. Duty-free shopping and a variety of recreational attractions, along with a refurbished beachfront, are among the primary attractions.

Finance. Singapore's financial services are highly sophisticated and are available through a wide variety of institutions. There is a growing venture-capital market that offers seed funding to firms that develop or introduce new technology. The government's Monetary Authority of Singapore performs all the functions of a central bank except issuing currency. A focal point of Singapore's growth as an international financial centre has been the Asian Dollar Market, which is essentially an international money and capital market where currencies other than the Singapore dollar are traded. The Development Bank of Singapore is the largest local bank in terms of assets. The Stock Exchange of Singapore is an important component of the financial activity in the region. The economic crisis that swept Asia in the late 1990s reduced the country's rapid economic growth, though Singapore weathered the crisis better than most of its neighbours.

Trade. Singapore continues to perform its traditional function as a financial intermediary, shipping raw materials such as rubber, timber, and spices from the Southeast Asian region in exchange for finished goods from both within and, especially, outside the region. Major imports are machinery and transport equipment and crude petroleum, while machinery, telecommunications products, and refined petroleum products are the major exports. The United States, Malaysia, and Japan are Singapore's principal trading partners. Entrepôt activities, where goods are transhipped and sometimes processed or manufactured in the immediate area, account for about one-third of Singapore's export trade. Notable in this capacity has been the oil-refining industry. In an attempt to foster additional trade, Singapore has become a joint-venture partner in numerous projects with Malaysia and Indonesia.

Entrepôt activities

The container terminal at Keppel Harbour, Singapore.
Manfred Gottschalk/Apa Photo Agency SIN

Transportation. Singapore has one of the world's busiest ports in terms of shipping tonnage. The Port of Singapore Authority oversees all shipping activity and operates a number of terminals on the island. Containerized cargo accounts for more than half of the general-cargo tonnage. The island has a well-developed network of roads and highways, but traffic congestion frequently is a serious problem. In the late 1980s and early 1990s the government opened a light-rail mass-transit system that links the major population centres in the housing estates with employment centres and the central business district. Singapore is linked by rail to Peninsular Malaysia via the connecting causeway at Johor. Singapore's international airport, Changi, at the eastern end of the main island, is a major regional and overseas air hub.

Administration and social conditions. *Government.* Singapore is a parliamentary democracy based on the Westminster model. The government consists of a president who is head of state and a unicameral Parliament of 81 members who are elected to terms of up to five years. The parliamentary majority selects the prime minister and cabinet from its own ranks, and they in turn form the government. Until 1991 the largely ceremonial post of president was filled by parliamentary election; in that year the constitution was amended to allow for the direct popular election of the president and for presidential powers to be expanded. In each constituency there is a Citizens' Consultative Committee, designed to link local communities to the ruling party.

Close liaison is maintained between the political and administrative arms of government. The administrative structure consists of the various ministries and statutory boards. These are staffed by civil servants who are monitored by an independent Public Service Commission.

The political process. Singapore's electorate includes every adult citizen who is a registered voter, and voting is compulsory. A number of parties contest elections, but since 1959 Singaporean politics have been dominated by the People's Action Party (PAP). The PAP's ability to maintain its control largely has been attributable to Singapore's rapid economic growth and improved social welfare. In addition, the PAP often has suppressed and co-opted domestic opposition—notably through internal-security laws that allow political dissidents to be held in-

The PAP

definitely without trial—and it has promoted a national paternalistic ideology through a variety of laws and corporate institutions. The emphasis of this ideology has been a rigid public morality focused on personal appearance and cleanliness, political loyalty, and family planning.

Justice. Justice is administered by the Supreme Court and by courts of lesser jurisdiction, such as district and magistrates' courts. Appeals can be made from the lower to the higher courts, with final appeal to the judicial committee of the Privy Council in London. A Sharī'ah court has jurisdiction in matters of Islāmic law.

Armed forces and security. The armed forces of Singapore are divided into army, air force, and navy branches. The army is by far the largest of the services and consists primarily of infantry battalions with supporting artillery, armour, engineer, and logistics units. The main duties of the air force are air defense, support of ground forces, and long-range surveillance and tracking. The navy patrols the country's coastal waters and protects shipping lanes. Compulsory military conscription for 18-year-old males was introduced in 1967. There are two paramilitary forces: the Peoples' Defence Force, composed mainly of reservists, and the National Cadet Corps, consisting of high-school and college students.

The police force is responsible for internal security, traffic management, and crime prevention. It is assisted by a Civil Defence Force consisting of reservists and volunteers.

Education. Education is highly valued in Singapore, and its education system is elaborately structured. Primary education is free and lasts from six to eight years; the language of instruction is English, and students are required to learn any one of the other three official languages as a second language. Students at the secondary level are placed into academic or vocational and commercial tracks. Those on academic tracks are further channeled into four- or five-year courses of instruction. Opportunities for higher education are determined by academic performance and usually involve two or three years of preuniversity instruction followed by enrollment at a university or technical college. The National University of Singapore, founded in 1980 by a merger of the University of Singapore and Nanyang University, is the largest and best-known institute of higher education.

Health and welfare. Health conditions in Singapore compare favourably with those in other economically developed nations. The range and quality of medical services is notably high, with a large number of doctors and dentists. There are both government and private hospitals, while nonhospital care is dispensed from numerous outpatient clinics and mobile centres. The government and voluntary associations, the latter coordinated by the Council of Social Service, provide welfare services for the aged, sick, and unemployed.

Cultural life. Cultural activities in Singapore are largely derivative, springing from one or another of the major civilizations of China, India, Indonesia, or the West. Traditional Chinese and Indian music, painting, and drama are practiced by numerous cultural societies and professional groups. Popular culture, based on modern mass media, is far more widespread. Malay music, which has adopted the rhythms of Western orchestras, has general appeal. Musical films that popularize Hindi and Tamil songs have a considerable following, as do films from Hong Kong, Taiwan, and the United States.

The press

Several Chinese, English, Indian, and Malay newspapers serve a largely literate population. Magazines published in the West, Hong Kong, and Japan also have wide appeal. The government monitors the press to a certain extent and on occasion places circulation restrictions on periodicals and newspapers that are critical of its policies. The government-owned Singapore Broadcasting Corporation controls all local radio and television broadcasting.

For statistical data on the land and people of Singapore, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Singapore Island originally was inhabited by fishermen and pirates, and it served as an outpost for the Sumatran empire of Śrīvijaya. In Javanese inscriptions and Chinese records dating to the end of the 14th century, the more common name of the island is Tumasik, or Temasek, from the Javanese word *tasek* ("sea"). Rājendra, ruler of the southern Indian Coḷa kingdom, attacked the island in 1025, and there was another Coḷa raid in 1068. In 1275 the Javanese king Kertanagara probably attacked Temasek when he raided Pahang on the east coast of the peninsula. According to a Chinese traveler, Wang Ta-yuan, just before 1349 about 70 Tai war boats besieged Temasek for a month but had to withdraw. The Javanese epic poem *Nāgarakeṛtāgama* (written 1365) includes Temasek among the conquests of the Javanese empire of Majapahit. At the end of the 14th century, Temasek fell into decay and was supplanted by Malacca (now Melaka). Yet in 1552 it was still a port of call from which St. Francis Xavier dispatched letters to Goa, and João de Barros described its busy shipping activity in his history *Décadas da Ásia* (1552–1615).

Rājendra may have named the city Singapura ("Lion City"), later corrupted to Singapore, or the name may have been bestowed in the 14th century by Buddhist monks, to whom the lion was a symbolic character. According to the *Sejarah Melayu,* a Malay chronicle, the city was founded by the Śrīvijayan prince Sri Tri Buana; he is said to have glimpsed a tiger, mistaken it for a lion, and thus called the settlement Singapura. (Rt.H./T.R.L.)

East India Company. In January 1819 Sir Thomas Stamford Raffles of the English East India Company, searching for a trading site, forestalled by the Dutch at Riau, and finding the Carimon (Karimun) Islands unsuitable, landed at Singapore. He found only a few Chinese planters, some aborigines, and a few Malays and was told by the hereditary chief, the *temenggong* (direct ancestor of the sultans of modern Johor), that the company could purchase land. The *temenggong,* however, was a subordinate of his cousin Abdul Rahman, sultan of Riau-Johor, who was under Dutch surveillance. Furthermore, Abdul Rahman was a younger son and not a sultan de jure. Raffles, disobeying instructions not to offend the Dutch, withdrew his own recognition of Abdul Rahman's suzerainty over Singapore and installed Abdul Rahman's elder brother, Hussein (Husain), to validate the purchase of land there on behalf of the company. The Dutch protested. In London the court of directors, though it decided Raffles had contravened instructions, took no action.

British rule

In 1824 an Anglo-Dutch treaty left Malaya and Singapore in the British sphere, and in August the whole of Singapore Island was ceded to the British for a monetary payment. Two years later Singapore, Penang, and Malacca (Melaka) were combined as the Straits Settlements to form an outlying residency of India. In 1830 they were reduced to a residency under Bengal, and two years later Singapore became their capital. When the East India Company lost its monopoly of the China trade (1833), it also lost its interest in Malaya. The settlements were transferred to the direct control of the governor-general of India in 1851. In 1867 they were made a crown colony under the Colonial Office in London.

Development of the port. Meanwhile, Singapore's trade had suffered after 1842 from British development of a rival port, Hong Kong, as later it was to suffer from the French occupation of the Indochinese Peninsula and their development of Saigon and Haiphong in Vietnam and from the establishment of Dutch ports and shipping lines in the Dutch East Indies. With the opening of the Suez Canal in 1869 and the advent of steamships, however, an era of prosperity began that led eventually to the construction of three miles of wharves at Tanjong Pagar and finally, in 1921, a naval base. The economic growth of the Malay states after they became British protectorates enlarged transit trade. (R.O.Wt.)

The demand of the industrial West for tin and rubber was what made Singapore one of the greatest ports in the world. After World War I, steps were taken to modernize Malayan defenses and, with the lapsing of the Anglo-Japanese alliance, to build a large naval base in Singapore.

World War II and the end of colonialism. In early December 1941 the Japanese landed in northern Malaya and

southern Thailand on the Malay Peninsula. They quickly gained air and naval superiority in the region, and by the end of January 1942 they had overrun the peninsula and were opposite Singapore Island. The Japanese crossed the Johor Strait on Feb. 8, 1942, and the British command surrendered the island and city one week later. Singapore remained in Japanese hands until September 1945.

Postwar British political plans for Malaya excluded Singapore from a proposed Malayan Union and later from the Federation of Malaya, mainly because it was thought that Singapore's predominantly Chinese population would be an ethnic obstacle to common citizenship. As a separate crown colony (from 1946), Singapore made constitutional progress despite the communist insurrection in Malaya. Elected ministers and a Legislative Assembly with an elected majority assumed government responsibility in 1955, except for matters of defense and foreign policy. In 1959 the official and nominated elements were eliminated, and Singapore became self-governing, although Britain still retained control of defense and foreign policy.

Singapore since 1963. Singapore joined the Federation of Malaysia on its formation in September 1963. The ruling People's Action Party (PAP), led by Lee Kuan Yew, had refused in 1959 to form a government until extreme left-wing leaders of the party who had been detained by the colonial authorities were released. These leaders opposed the concept of Malaysia and broke away from the PAP to form the Socialist Front (Barisan Sosialis), which was accused of being a communist front organization. The PAP faced fresh dangers of subversion when Indonesian opposition to Malaysia took the form of military and economic confrontation (1964).

Secession from Malaysia

Confrontation ended in 1966, but Singapore had seceded from Malaysia in 1965 (at the invitation of the Malaysian government) because of political friction between the state and central governments. This conflict had ethnic overtones and continued to affect relations between Singapore and Malaysia until the mid-1970s, when relations became more cordial.

In January 1968 the British government had announced that all British defense forces would be withdrawn from East and Southeast Asia (except Hong Kong) by the end of 1971. In April Singapore's unprepared major opposition parties boycotted an election called seven months before it was due. The ruling PAP termed its sweep of all parliamentary seats a mandate for its plans for reducing the economic effects of the British military withdrawal.

At the end of October 1971, British military presence in Singapore came to an end. The Anglo-Malayan treaty concluded in 1957, which had committed Britain to the defense of the region, was terminated, and in its place a five-power defense arrangement—involving Britain, Australia, New Zealand, Malaysia, and Singapore as equal partners—came into force.

Since the 1970s Singapore has pursued an aggressive policy of economic growth based primarily on export manufacturing and trade. Gradually, it also has taken a more active role in regional diplomacy. Singapore was a founding member of the Association of Southeast Asian Nations (ASEAN) in 1967, and by 1980 it had emerged as one of ASEAN's leaders. The PAP has continued to dominate Singaporean politics, although Lee stepped down as prime minister in 1990, and between 1981 and 1991 opposition parties gradually increased their number of seats in Parliament from one to four. Yet, despite the country's phenomenal economic success, resultant high standards of living, and subsequent goal of internationalization, the government's policies of developmental paternalism have bred some discontent among those who have come to expect greater openness to new ideas and a freer flow of information. (A.Ke./T.R.L.)

For later developments in the history of Singapore, see the BRITANNICA BOOK OF THE YEAR.

Thailand

Located in the centre of mainland Southeast Asia, the Kingdom of Thailand (Thai: Muang Thai, or Prathet Thai) has emerged as a pivotal nation in the region's political and economic life. The country was officially called Siam until 1939 and again briefly in 1945–48. The several ethnic and religious groups represented among Thailand's people are characteristic of the cultural diversity that for centuries has spread southward from China and eastward from India. Indeed, the name "Thai" to describe the country's people came into use only in the 20th century.

Thailand's area of 198,115 square miles (513,115 square kilometres) consists of two broad geographic areas: a larger main section in the north and a smaller peninsular section in the south. The main body of the country is surrounded by Myanmar (Burma) to the west, Laos to the north and east, Cambodia to the southeast, and the Gulf of Thailand (Gulf of Siam) to the south. Peninsular Thailand stretches southward from the southwestern corner of the main part down the Malay Peninsula; Myanmar extends along the western portion of the peninsula as far as the Isthmus of Kra, after which Thailand occupies the entire peninsula until reaching its southern border with Malaysia at roughly latitude 6° N. Bangkok, Thailand's capital and chief port, is in the main portion at the head of the Gulf of Thailand.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Thailand's landscapes vary from high mountains to fertile alluvial plains dotted with rice paddies to sandy beaches set amid the equatorial latitudes of the Asian monsoons. Three physiographic regions cover most of the country: the folded mountains in the north and west, the Khorat Plateau in the east, and the west-central Chao Phraya River basin. The maritime southeastern corner of the main portion and the long, slender peninsular portion in the southwest constitute separate physical regions.

Relief. The northern mountains, the southeastern continuation of the uplift process that formed the Himalayas, extend southward along the Thai-Myanmar border and reach as far south as northern Malaysia. Long granitic ridges were formed when great masses of molten rock forced their way upward through the older sedimentary strata. Peaks average about 5,200 feet (1,585 metres) above sea level. Mount Inthanon, at 8,481 feet (2,585 metres) the highest in the country, is in northwestern Thailand, near the ancient city of Chiang Mai (Chiengmai); the city is overshadowed by Mount Suthep, a tourist attraction and site of the royal summer palace. Some of the rugged limestone hills contain caves from which remains of prehistoric humans have been excavated.

The Khorat Plateau

The Khorat Plateau is a vast tableland bounded by the Mekong River on the north and east. It was formed by uplifting along two perpendicularly arranged crustal faults—one trending north-south in the west and the other east-west in the south. As a result, the underlying sedimentary rocks were tilted rather than uniformly uplifted. This tilting created ranges of low hills and mountains along the western and southern edges of the plateau: the Phetchabun and Dangrek (Thai: Dong Rak) mountains, respectively. The escarpments of these uplands overlook the plain of the Chao Phraya basin to the west and the Cambodian plain to the south. Surface elevations on the Khorat Plateau range from about 650 feet in the northwest to about 300 feet in the southeast: the terrain is rolling, and the hilltops generally slope to the southeast in conformity with the tilt of the land.

Situated between the northern and western mountain ranges and the Khorat Plateau is the extensive Chao Phraya River basin, which is the cultural and economic heartland of Thailand. The region, sometimes called the Central Plain, consists of two portions: heavily dissected, rolling plains in the north and the flat, low-lying floodplain and delta of the Chao Phraya in the south. It was formed by the outwash of immense quantities of sediment brought down from the mountains by the Chao Phraya's tributaries, which produced vast fan-shaped alluvial deposits.

The generally rolling countryside of the southeast has high hills in the centre and along the eastern boundary with Cambodia. Notable peaks are Mount Khieo (2,614 feet), visible from the top of Mount Phu Thong (the Golden Mount) in Bangkok, and Mount Soi Dao, which attains a height of 5,471 feet. The hills, reaching nearly to the sea, create a markedly indented coastline fringed with many

islands. With their long stretches of sandy beach, such coastal towns as Chon Buri and Rayong and some of the islands have become popular year-round tourist resorts.

The topography of the peninsula is rolling to mountainous, with little flat land. A gently sloping sandy coastline, including such resort communities as Hua Hin, borders the Gulf of Thailand on the east. Higher mountains reaching about 4,900 feet line the peninsula on the west and contain narrow passes linking Thailand and Myanmar. These ranges separate the Andaman and South China seas as the peninsula narrows near the Malaysian border. Off the rugged and much indented west coast lie numerous large islands, including tin-rich Phuket Island.

Drainage and soils. Thailand is drained largely by two river systems: the Chao Phraya in the west and the Mekong in the east. Three major rivers in the northern mountains—from west to east, the Ping and its tributary the Wang, the Yom, and the Nan—flow generally south through narrow valleys and onto the plains and then merge to form the Chao Phraya, Thailand's major river.

The Chao Phraya delta

The delta floodplain of the Chao Phraya is braided into numerous small channels and is joined by other rivers—notably the Pa Sak—as the river flows toward its mouth in the Gulf of Thailand. The flooding of the flat delta in the wet season is an asset to rice cultivation, although higher ground on the extreme eastern and western edges of the plain requires irrigation. The entire delta was once part of the Gulf of Thailand, but over time the sediments carried down from the north have filled it in. Such silting is a continuing obstruction to river navigation, but it also extends the river's mouth into the gulf by several feet each year.

The rivers of the Khorat Plateau flow generally southeastward and empty into the Mekong. Monsoon rains over the degraded forest cover of the region produce rapid runoff; flooding occurs almost yearly at Ubon Ratchathani at the junction of the plateau's two major rivers, the Mun and the Chi. Swampy land and lakes are common close to the Mekong, in contrast to the aridity of much of the rest of the region. A high groundwater table there contains mostly brackish, unpotable water. The Mekong itself is either studded with islands or broken up by impassable rapids.

The southeast and the peninsula are drained by short streams and rivers. In the southeast the rivers in the north flow into the Chao Phraya delta, while those in the west and south run directly into the sea, where they have built up small alluvial basins and deltas along the coast; the mouths of the rivers along the southern coast consist of tidal flats and mangrove swamps. Nearly all the rivers on the peninsula drain into the Gulf of Thailand. The Phet River is one of several that has been dammed for irrigation.

The great deposits of alluvial soils in the river valleys are the most fertile in Thailand. These are replenished annually with sediment washed down by rivers swollen with the annual monsoon rains. Chief among these areas is the delta floodplain of the Chao Phraya, but the relatively flat basins in the northern mountains, scattered lands along the Mun and Chi rivers on the Khorat Plateau, and much of the coast also have rich alluvial soils. Soils elsewhere tend to be relatively infertile, highly leached laterites. Near the Mekong, a high salt content in some soils limits crop production, although salt deposits there are mined commercially.

Climate. The major influences on Thailand's climate are its location in the tropical monsoon zone of mainland Southeast Asia and certain topographic features that affect the distribution of precipitation. Beginning in May, the warm, humid air masses of the southwest monsoon flow northeastward over the region from the Indian Ocean, depositing great quantities of precipitation; rainfall reaches a maximum in September. The wind pattern is reversed between November and February, when the northeast monsoon brings cool, relatively dry air masses in a southwesterly flow to create a seasonably cooler climate for much of the country. Stagnant air in March and April is associated with a distinct hot and dry intermonsoonal period.

Local relief acts to modify these general weather patterns. Topographic effects are most noticeable on the peninsula, where Ranong on the west coast receives approximately 160 inches (4,000 millimetres) of precipitation annually, while Hua Hin on the east coast receives less than 40 inches. Similar but less pronounced rain-shadow effects occur along the western margins of the Central Plain and on the Khorat Plateau. Songkhla, at the southern end of peninsular Thailand, has its rainy season during the cool season, the result of moisture picked up by the northeast monsoon winds while passing over the Gulf of Thailand; in this area a true tropical rain-forest climate prevails.

Temperatures

Nationwide, temperatures are relatively steady throughout the year, averaging between 77° and 84° F (25° and 29° C). The greatest fluctuations are in the north, where frost may occur in December at higher elevations; conversely, maritime influences moderate the climate in the south. The cooler, drier air of the northeast monsoon produces frequent morning fogs that generally dissipate by midday in the north and northeast regions.

Plant and animal life. Since 1970 the area of Thailand that is covered with forest has dwindled from about half to one-fourth. Forest clearing for agriculture, excessive logging, and poor management are among the main causes of this decline. Less than one-fifth of the country is covered by grass, shrubs, or swamps, and the remainder is under settlement or cultivation. Forests consist largely of such hardwoods as teak and timber- and resin-producing trees of the Dipterocarpaceae family. As elsewhere in Southeast Asia, bamboo, palms, rattan, and many kinds of ferns are common. Where forests have been logged and not replanted, a secondary growth of grasses and shrubs has sprung up that often limits land use for farming. Lotuses and water lilies dot most ponds and swamps throughout the country.

Elephants, buffalo, cattle, horses, and mules are among the domesticated animals important for agriculture and transportation. Only a few elephants are still found in the wild, and, although forestry is now largely mechanized, elephants remain helpful in difficult terrain. Mechanized agriculture is also rapidly replacing draft animals, and horses and mules increasingly are used only as pack animals along the mountain trails of the north. The forests abound in monkeys and species of birds. Large mammals such as the rhinoceros and tapir, however, are among the species endangered by deforestation and game hunting. The government has created a number of wildlife and conservation areas to protect these valuable resources and has acted to prevent the illegal sale of endangered species.

Lizards live in settled areas and prey on insects. Frogs and toads (some of them edible) are numerous, crocodiles are common in the south, and snakes are abundant, especially the king cobra and various species of dangerous water snakes. Several farms in Bangkok raise crocodiles and snakes for commercial and scientific purposes.

The once-abundant freshwater and marine fish have been rapidly depleted by overfishing and disruption of their natural habitats, as have such edible crustaceans as shrimp, prawns, and sea crabs. There are many species of wild silkworms, and some are raised for the silk industry. Mosquitoes still transmit malaria, although the incidence of this disease has been decreasing. White ants and moths are a scourge to clothing and books.

Traditional regions. Thailand can be divided into five historical regions: Lanna Thai (northern Thailand), Isan (northeastern Thailand), the Central Plain, Pak Tai (southern Thailand), and west and southwest Thailand.

The people of the mountainous Lanna Thai speak a dialect of the Thai language called Kham Mu'ang, or Yuan in its written form, and follow Buddhist traditions more akin to those practiced in Myanmar. They also share a preference with the Lao-speaking Thai of northeastern Thailand and the people of Laos for glutinous rice as their staple food. The mountains of the north also are the home of many upland minority groups that have migrated from Myanmar, Laos, and southern China over the centuries.

The Isan region shares various linguistic, artistic, and religious traditions with Lanna Thai and the Lao across the Mekong River. The regional dialect is referred to as Lao or Isan, but most people can easily communicate in standard Thai. This region also had a history of relative

autonomy until the late 18th century, which has helped to shape northeastern Thai identity.

The Central Plain

The Central Plain, occupying most of the Chao Phraya basin, is the core cultural region, and its people (often called the Siamese) speak the national language—standard Thai, or Siamese. Historically, the Siamese followed a Buddhist tradition that has been more closely linked to that of the Khmer of Cambodia. Because of the plain's central position, population and economic activity are heavily concentrated there, especially in Bangkok, and industrial and commercial enterprises have grown faster there than elsewhere. This rapidly growing economic heartland continues to be a strong magnet, attracting people from the northeast and other places who are seeking greater economic and social opportunities.

The southeast, lying close to the sea, is an undulating and hilly region extending eastward from Bangkok to the Cambodian border. Sino-Thai, or Thai of Chinese descent, are a prominent segment of the regional population, their ancestors having originally settled there in the late 19th century to work on sugarcane plantations, in lumber mills, and as small merchants. There is also a significant number of people living along the border with Cambodia who speak Khmer or Khmer-related languages and follow distinctive traditions.

The southern Pak Thai region has a distinctive identity linked to the historical role of such towns as Nakhon Si Thammarat, once called Ligor. Because this background is related to the later Siamese kingdoms, the language and customs of Pak Thai are similar to those of the Siamese of central Thailand. The extreme south is inhabited by Malay-speaking Thai, most of whom are Muslim.

The west and southwest, consisting mostly of hilly to mountainous terrain adjoining the Myanmar border, are sparsely populated. The Karen from Myanmar often migrate and live within Thailand, clearing and cultivating uplands in a manner similar to the upland minorities of northern Thailand.

Settlement patterns. Hill settlements depend much on shifting cultivation of upland crops. Such mountain peoples as the Karen, Hmong (Miao, or Meo), Yao, Lahu, Lisu, and the lowland Thai who have migrated to the uplands usually settle on the ridges and the slopes in groups ranging up to 100 or more houses, depending on the resources of the area. The Hmong are opium cultivators and prefer to live on high slopes where opium grows well in the cool climate. The Karen live along the stream valleys and grow rice on well-tended terraced fields. The lowland Thai who have migrated to these uplands earn a living from their tea and coffee plantations.

No true plains villages exist in Thailand. Villages in the northeast are scattered on the higher grounds above flood levels, while the lower grounds are used for rice farming. In the north, where the villages are found in the alluvial basins of major rivers, increased population and transportation have tended to disperse the villages away from the rivers and toward the main railroads and highways, reducing the amount of land available for growing rice.

The Chao Phraya delta is densely settled but only on the high ground that is free from flooding. A vast network of irrigation canals modifies the pattern of settlement. With increasing facility in transportation offered by small motorboats, the villages tend to become dispersed to the east and west away from the rivers. New highways also tend to modify settlement patterns, especially at their crossings of canals and rivers, where new towns grow up rapidly.

In the south and southeast, plantations, especially fruit and rubber plantations, are scattered along the fertile slopes, alternating with the low and narrow rice fields; the villages are therefore arranged accordingly. Most of them are joined by good roads and highways. Alluvial deposits containing tin, no matter how remote, are accessible by land and sea. Settlement is almost continuous along both sides of the peninsula. Most of the people live by fishing, except in areas where collecting bird's nests for cooking brings a good income. The coastal villages are connected by both land and sea.

Urbanization

Urban settlement in Thailand, as in many other developing countries, has grown dramatically since World War II,

A Lisu hill settlement near Pai, northwestern Thailand.
Jeffrey Alford/Asia Access

but that growth has been highly uneven. Bangkok remains the dominant and only major urban centre in the country, its population more than 20 times larger than that of Nonthaburi, the next largest city. The smaller urban centres typically are provincial capitals; among the oldest of these are Chiang Mai, Lampang, Phrae, and Nan in the north, which grew up along the major tributaries of the Chao Phraya River.

The people. The diversity of ethnic, linguistic, and religious groups in Thailand is characteristic of most nations of Southeast Asia, where shifting political boundaries have done little to impede the centuries-long migrations of people. Thailand's central position on the mainland has made it a crossroads for these population movements. As a result, speakers of all four major language families are represented in the country.

Ethnolinguistic composition. People speaking Tai languages constitute by far the dominant linguistic group in the country. The largest group of Tai speakers are the Thai, who constitute more than half the population. The Thai live in almost all areas of the country and speak related dialects that are differentiated by accents and a few words. The most common dialect is called standard Thai (Siamese), with the greatest concentration of speakers in the Chao Phraya delta. Since standard Thai is the national language, which is used in all schools and official publications and by the national press and broadcast media, most Tai speakers can communicate adequately among themselves.

Tai-speaking peoples are found not only in Thailand but also in Myanmar, Laos, Vietnam, and southern China. Little difference exists among the Tai speakers in Laos, Myanmar, China, and northern Thailand, but there is a noticeable difference between these peoples and the Thai living in the Central Plain and close to Cambodia. Speakers of Lao are the largest linguistic minority in Thailand; they live in the Khorat Plateau adjacent to Laos and constitute about one-fourth of Thailand's population.

Wars pitting Thailand against Myanmar and Cambodia in the past brought many refugees and prisoners of war into Thailand. The Mon, a people of Myanmar, settled in many parts of the north, centre, and west, although they are now concentrated in an area just west of Bangkok. The Khmer settled in the east along the Cambodian border. Both groups traditionally spoke languages of the Mon-Khmer group of the Austro-Asiatic family; most of these people, however, now use standard Thai, many speaking it as a first language.

Among speakers from the other two language families, Malays at the southern tip of peninsular Thailand are the most numerous. The Karen and other speakers of Sino-

Tibetan languages sparsely inhabit the western and northern mountains. Except for the Karen, who mixed rather easily with the northern Thai, the hill tribes prefer to keep themselves isolated. They occasionally come down to the markets to trade with the lowlanders. Two small hill tribes, the Lawa (or Wa) and Semang, are of special interest. The Lawa, who speak a Mon-Khmer language, are believed by some historians to be the original dwellers of the delta plain, who subsequently were driven into the hills of the northwest by the Tai speakers who conquered the area. The Semang of the southern mountains speak a dialect of Malay and live by hunting with blowpipes and spears.

The Lawa and Semang

The Chinese constitute a significant minority in Thailand. In the commercial centres of Bangkok and other cities, people of Chinese descent operate both large and small commercial enterprises. The Chinese also make a living as middlemen and storekeepers. Most of them speak Chinese, although many also speak standard Thai.

English is also widely used in Thailand, especially in the urban centres. English is a required subject in secondary schools and the universities, and frequent contact with English-speaking foreigners also encourages the use of English. The prevalence of various South Asian dialects reflects the large number of Indian merchants and their descendants in the commercial centres.

Religions. Buddhism is professed by the vast majority of Thailand's population and is considered the national religion; most Thai Buddhists are of the Theravāda school. Muslims constitute a sizable minority and live mostly in the south. Most of the country's small Christian community lives in the central region. Hindus also are concentrated in the central region, chiefly around Bangkok. Although several of the hill tribes have converted to Buddhism or Christianity, most remain animists.

An unusual aspect of Thai religious life is the considerable influence of the Hindu Brahmans, even though they total only a few thousand families. Most royal and official ceremonies are almost always directed or performed by the Brahmans, whose rites are blended harmoniously with those of the Buddhists. Brahmans are renowned for their astrological expertise, assume responsibility for preparing the national calendar, and officiate at such state ceremonies as the annual plowing ceremony, which is believed to bring a good rice harvest. There are no important new religious movements in Thailand, but Buddhist monks have become more vocal in advocating environmental and social issues

Demographic trends. Thailand's population has increased rapidly in the 20th century, especially between 1950 and 1970, when the government supported such growth. Since then, official policies and private family-planning programs have slowed this growth dramatically, making the country a model for other nations seeking to reduce their high growth rates. The more youthful age profile of the population that resulted from the earlier growth, however, has begun to place increasing demands on the country's education, housing, health, and employment systems.

Population movements

Internal migration in Thailand occurs primarily between rural areas and has had little effect on the regional distribution of population. Nonetheless, migration from the countryside has contributed significantly to the growth of Bangkok. Since 1960 Bangkok has received two-fifths of all interregional migrants in the country, many of whom have come from the central and northeast regions. As in most other areas of the world, these migrants are mainly young adults less than 30 years of age. After 1975, Thailand also received large numbers of Cambodian, Vietnamese, Laotian, and Hmong refugees who fled political conflicts in their own countries and settled in camps along the Thai border; most refugees, however, have been resettled in other countries or repatriated to their own countries.

The economy. Thailand's investment-oriented economy is among the most rapidly growing in Asia. Despite this success, economic development has been highly uneven, especially in agriculture. Although much of Thailand's export revenues and a majority of the labour force depend on agriculture, its contributions to economic growth have declined consistently since 1950. Aiming at diversification, the government has encouraged investment in small industry. To encourage exports, duties are low, except on rice, to which a premium is attached to prevent domestic shortages. Unions are prohibited and strikes not allowed unless management fails to agree with employees and government mediators.

Resources. Tin, mined mostly in the peninsula, long has been among Thailand's most valuable mineral resources, and the country has become one of the world's largest producers. The construction of a smelter made it possible to process most of the ore domestically. Fluctuations in the world tin market, however, have caused production to be reduced. Other important mining and quarrying operations produce coal (lignite), zinc, gypsum, fluorite, tungsten, limestone, and marble. Rubies and sapphires are mined along the east coast of the peninsula. Thailand is one of the world's largest exporters of gems and jewelry, and these are among the country's top sources of foreign exchange.

Agriculture, forestry, and fishing. Growing international market demand has contributed to a more diversified Thai agriculture. Rice is the main staple food crop, but foreign demand for other crops has significantly expanded the production of cassava, corn (maize), kenaf (a fibre used as a substitute for jute), and sugarcane; rice yields, however, are among the lowest in Asia. Despite the Thai farmers' interest in growing crops to meet this demand, the government has done little to help farmers withstand sharp fluctuations in the prices of major export crops.

The main agricultural region of Thailand, the Chao Phraya basin, is planted mostly in rice. Rice also is the principal crop of the cultivated regions of the Khorat Plateau. In the northern intermontane basins, rice, vegetables, tobacco, and fruit trees are raised. Large inland areas of the southeast and peninsula contain cassava, sugarcane, pineapple, and rubber plantations. Cattle breeding has been curtailed by disease among stocks, but pigs and poultry are widely raised, with the latter constituting a growing export commodity.

Hardwoods such as teak and yang (a source of gurjun balsam), once major forest products, are now exported only in small amounts, especially since a government ban on logging imposed in 1989. Rubber production—based on trees introduced into the country during the 19th century—is important in the south. Fish production includes both marine fishes and freshwater species caught in rivers and reservoirs or raised in ponds.

Rubber

Industry. The growth in manufacturing since 1970 has been especially dramatic, reflecting the large investments made by private firms. Clothing, canned goods, and electrical circuits are among the more important products exported. Many larger concerns have been financed by foreign and Thai capital. Japan, South Korea, Taiwan, and Singapore have been major sources of investment for new industries, especially those producing goods for local use. Most industry and manufacturing is concentrated in or near the Bangkok metropolitan area, while numerous cottage industries in the north produce textiles, teak carvings, lacquerware, and other craft products. In the southeast, oil refining and gem cutting, in addition to food processing, are carried on.

Thailand has several hydroelectric plants, but most electric power is produced by generating plants using natural gas and lignite. Only small amounts of petroleum are produced domestically; and most of the oil consumed in the country must be imported. Somewhat larger quantities of natural gas are recovered from the Gulf of Thailand and the northeast and contribute to national energy needs.

Finance and trade. The Bank of Thailand, established in 1942, issues currency, acts as central banker to the government and to the commercial banks, and serves as fiscal agent in dealing with international monetary organizations. More than half of the nation's retailing and other distribution businesses are located in the Bangkok metropolitan area. Middlemen handle most farm commodities. Retail stores are small, except in Bangkok, which has many large department stores and shopping malls. Prices in these stores are fixed, but bargaining occurs regularly in smaller shops and in markets throughout the country.

With agricultural and raw materials as the basic exports, manufactured goods, such as machinery and transportation equipment, account for the highest value among imports. Thailand's major trading partner is Japan; other important partners are the United States, Singapore, Germany, and Hong Kong.

Transportation. Bangkok is the centre of Thailand's water, land, and air transport systems. The rivers of the Chao Phraya delta have been used since antiquity, and modern irrigation canals have added to the inland-waterway network. Thailand has undertaken substantial highway development, but the seasonal rains make it difficult to keep some roads open year-round, especially in the peninsula. Mountain trails are often the only means of travel in remote areas. Where roads are inadequate, airplane and helicopter services often compensate. Rail lines radiate from Bangkok, with one linking up with the Cambodian rail system. The port of Bangkok, at Khlong Toei, is the largest and busiest in the country, handling nearly all imports and exports. Don Muang International Airport, north of Bangkok, is the hub of Thailand's air network; more than 20 smaller airports are located throughout the country.

Administration and social conditions. *Government.* Following a coup in 1932 that ended the absolute rule of the monarchy, a constitution was promulgated, in which the monarch, the National Assembly, the State Council, and the law courts were to exercise power on the behalf of the citizenry. Since then, several constitutions have been created because of changes of government, but the provisions have remained similar to the 1932 document. Under the present constitution, the king is head of state and of the armed forces. He is held to be sacred and inviolable, and in the name of the people he exercises legislative power, with the advice and consent of the National Assembly. He also appoints the prime minister. Executive powers are vested in the prime minister and cabinet, who operate in the name of the king. The royal family is very much at the core of modern Thai society, being regarded as the symbol of national unity and the protector of national welfare and traditions.

The 1932 constitution

In form, the Thai government resembles those of Western nations: various ministries are responsible for such matters as finance, agriculture, education, public health, communications, and justice. Despite this similarity, new administrations frequently have come to power through military-backed coups. These authoritarian governments dominated Thai politics until 1973, when a student-led movement forced a return to constitutional rule, popular elections, and registration of political parties. Since then, there has been a succession of military-installed or popularly elected governments in Thailand. Stronger democratic principles and opposition to military rule throughout the country make any future return to more authoritarian governments unlikely.

The provinces (*changwat*), of which there are 73, are the major units of local government. Below these are districts, subdistricts, communes, and villages. Municipalities in the kingdom are classified as cities, towns, or communes, according to their populations; they are run by an elected mayor and councillors.

Justice. Thai law has been influenced by the Hindu code of *Manu-smṛti,* which probably was transmitted through the ancient Mon kingdom of central Thailand. Reform in the late 19th century introduced concepts of Western jurisprudence. All judges in the country's courts are professionals, appointed without political consideration; they are bolstered by a system of judge trainees.

Armed forces. Under the king as commander in chief, the army, navy, and air force are assisted by the United States Military Assistance Program. Thai soldiers have fought in Korea and Vietnam, but since 1975 the military has been more fully occupied with protecting its own borders from problems in the neighbouring countries of Cambodia, Laos, and Myanmar.

Education. Children under 15 are required to complete six years of elementary education. Secondary education generally lasts six years and consists of lower and upper divisions, each lasting three years; students at the upper-secondary level have the choice of pursuing an academic or vocational program. Only a small minority of students, however, go on to secondary training. Ramkhamhaeng University in Bangkok, established as an open-enrollment university in 1971, has considerably increased opportunities for advanced education in the country. In addition, several other institutions, not including military academies, offer degrees in undergraduate and postgraduate fields. The oldest and largest of these is Chulalongkorn University in Bangkok, established in 1917.

Universities

Health and welfare. Thailand's health and social-welfare services remain far from adequate. The emigration of potential medical practitioners to more lucrative practices in the West has tended to undermine the government's attempts to upgrade services within the country. Mobile medical centres and helicopters attempt to alleviate the concentration of facilities in Bangkok and the regional centres. The doctor-patient ratio is poor, and medical practice on the midwife level is common. Infant mortality and diseases of childbirth are leading causes of death, whereas malaria has been widely reduced through the use of pesticides, although such agents as DDT also have contributed to the devastation of many forest animals as a side effect.

Only Bangkok and a few other large municipalities have housing shortages. The construction of government-financed housing cannot keep pace with demand, and slum areas occupy some parts of the city.

Cultural life. It formerly was thought that the Thai's original home was in China, but it is now generally believed that Tai-speaking peoples originated in northern Vietnam and began settling the Indochinese Peninsula and southern and southwestern China about 1,000 years ago. These people, however, brought with them many cultural institutions of the Chinese. As they moved southward into what is now modern Thailand, they encountered Mon and Burman peoples from the west, Javanese from the south, and Khmer from the east and were influenced by their cultural traditions. In addition, an Indian presence already had been established in the region, and the continuous absorption of Indian culture became a significant component of Thai cultural development.

The royal palace plays an important role in leading and preserving Thai culture through frequent royal functions and state ceremonies. Among these is the *kaṭhina,* or robe-offering, ceremony, a colourful pageant marking the end of *vassa,* the period of Buddhist monastic retreat. It takes place with a procession of royal barges on the Chao Phraya River, reconstructing a tradition dating from the earliest days of Buddhism. Thai temples hold ceremonies to mark the special events of the Buddha's life, which often are accompanied by fairs attracting large crowds to the temples.

Silpakorn University in Bangkok provides training in all Thai fine arts, including drama and music. The university also designs buildings for the government and for religious institutions in styles that preserve traditional Thai architectural forms. Other important national cultural institutions include the Royal Institute, the Siam Society, and the National Museum, all located in Bangkok.

The arts. Religion has had a major influence on Thai artistic expression and is especially manifested in the sculpture of Buddhist images. Traditional Thai architectural style also is best seen in the multiple-structured temple compounds. Wood is usually the basic construction material, with the walls made of bricks and plaster. The ornamental parts are generally gilded and enriched with glass mosaic, gold leaf, porcelain, stucco, lacquer, and inlaid mother-of-pearl. Remnants of the original sites of palaces and temples can still be found in many of the old provincial centres. In Chiang Mai the old, square city walls are still extant, with numerous Buddhist temples scattered inside and outside the walls.

Architecture

Porcelain and pottery, at first made primarily for utilitarian purposes, later came to be regarded and fashioned as objects of art. Thai painting is probably derived from India and Sri Lanka and is mostly religious. The paintings, executed by anonymous monks or dedicated laypeople, are usually drawn on temple walls.

Traditional Thai music shares close affinities with the musical forms of Laos and Cambodia. There are three orchestral types: *pi phat,* which are used at court ceremonies and in the theatre; *kruang sai,* which perform at village festivals; and *mahori,* which accompany vocalists.

Press and broadcasting. The first type for printing the Thai written language was derived by a British military officer in 1828, and the first printing press was brought to Thailand by an American missionary in 1836. The Thai government first made use of the printing press in 1839, when a royal proclamation banning opium smoking and trade was printed. The press is the oldest of the mass media, with daily newspapers in Thai, Chinese, and English. The state directly controls radio and television broadcasting, but newspapers and periodicals are under private ownership and enjoy a wider freedom than elsewhere in Southeast Asia. Great care is taken, however, to depict the royal family and monarchy positively. (P.P.A./Ja.A.H.)

For statistical data on the land and people of Thailand, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The modern Thai are descended from a much larger group of peoples who speak Tai languages. Tai-speaking peoples are found from extreme northeastern India in the west to northern Vietnam in the east and as far south as the central Malay Peninsula. In the past scholars held that a parent group called the Proto-Tai originated in southern China and pushed south and west from the China landmass into northern mainland Southeast Asia. Most scholars now believe that the Tai came from northern Vietnam around the Dien Bien Phu area and that about 1,000 years ago they spread from there northward into southern China, westward into southwestern China, northern Myanmar (Burma) and northeastern India, and southward into what are now Laos and Thailand.

Early Tai culture. The Tai were lowland peoples who historically settled along river valleys in northern mainland Southeast Asia and southwestern China. There they formed small settlements where they practiced subsistence agriculture based on rice cultivation, supplemented by fishing and gathering forest products. Early in their history the Tai domesticated animals: they used water buffalo for plowing and ritual purposes and pigs and fowl for food. Women were accorded relatively high social status and could inherit property. The Tai practiced animism; they believed that spirits could be benevolent or malevolent and needed to be propitiated through offerings and special ceremonies.

The *müang*

The basic unit of Tai political organization was the *müang,* or group of villages, ruled by a *chao,* or hereditary chief or lord. During the 1st millennium AD the political strengths of the *müang* system enabled the Tai to move out of their original homeland until, by the 8th century, they had expanded across much of northern mainland Southeast Asia. By the 11th century they had begun to filter down into the area of present-day Thailand, and by the middle of the following century they had formed petty principalities there.

Mon-Khmer civilizations. As the Tai moved into mainland Southeast Asia, they came in contact with peoples speaking Mon-Khmer languages who had long inhabited the region. During the early centuries of the 1st millennium AD, Indian traders traveling to China had carried Hindu and Buddhist beliefs and practices to some of these peoples. Among them were the Mon of Myanmar, the first peoples in mainland Southeast Asia to adopt Buddhism. Between the 6th and 9th centuries the Mon established several small Buddhist kingdoms within the area encompassed by present-day southern Myanmar and central Thailand. From what are now the towns of Nakhon Pathom and Lop Buri (Lopburi) in west-central Thailand, they extended their power eastward across the Khorat Plateau, northward as far as Chiang Mai (Chiengmai), and northeastward into what is now Laos. These Mon kingdoms collectively are called Dvaravati. The Dvaravati period is noted for its artwork, particularly its Buddhist sculptures and votive images made of terra-cotta or stucco.

As the Tai moved south into mainland Southeast Asia, they also encountered the Khmer of Cambodia. Between the 9th and 13th centuries Khmer rulers expanded their domains from their capital at Angkor, establishing an empire that at its height under Jayavarman VII (ruled 1181–*c.* 1220) extended over approximately half of modern Thailand. While Mon kingdoms were predominantly Buddhist in character, Khmer civilization—which found its supreme expression in the great temple complex at Angkor—was heavily influenced by Hindu ideas and practices. Tai contacts with the Khmer led to many Hindu elements entering Tai culture, particularly in regard to royal ceremonies or classical dance and literature. Many of these elements can still be found in modern Thai culture today.

By the beginning of the 13th century, the Tai were starting to place pressure on both the Mon and Khmer empires. The Tai were settled throughout the Chao Phraya basin, and a Tai ruler was established as far south as the principality of Nakhon Si Thammarat, on the Malay Peninsula. Through Nakhon Si Thammarat a new form of Buddhism—Theravāda—had entered mainland Southeast Asia from Sri Lanka. Theravāda Buddhism was carried by monks not only to areas under Mon or Khmer rule but also to the new Tai principalities that were beginning to emerge. Sukhothai and Lan Na (Lanna), the first major Tai kingdoms in Thai history, were Theravāda Buddhist.

Sukhothai and Lan Na. The kingdom of Sukhothai, situated in the upper Chao Phraya basin, was founded in the mid-13th century when a local Tai ruler led a revolt against Khmer rule at an outpost of the Khmer empire. During its first two reigns, Sukhothai remained only a small local power. Under its third ruler, Ramkhamhaeng (ruled *c.* 1279–98), however, Sukhothai power was extended to the south as far as Nakhon Si Thammarat, to the west into what is now Myanmar, and to the northeast as far as Luang Prabang in modern Laos. Not all these territories were conquered by force: many became vassal or tributary states to Sukhothai based on ties of kinship or personal loyalty and linked to it in a loose confederation.

Ramkhamhaeng

Ramkhamhaeng is renowned not only for extending the territory under Sukhothai control but also for leaving a remarkable stone inscription, which is considered by most scholars to contain the earliest example of writing in the Thai language. Written in 1292 and utilizing Khmer script adapted to the sounds and tones of Tai speech, it pictures the Sukhothai kingdom as prosperous, active in trade, and benevolently governed by a paternal monarch. According to the inscription, the state taxed its citizens modestly, treated all citizens (including non-Tai) alike, and provided justice for all. The Sukhothai period (mid-13th to mid-15th century) also is noted for its sculpture and pottery. Graceful bronze sculptures of the Buddha, especially those showing him in the walking position, are typical of the period, while the celadon ware made at Sukhothai and nearby Sawankhalok was exported throughout Southeast Asia.

Sukhothai was not the only Tai state in Southeast Asia during this period. In the mid-13th century in what is today northern Thailand, a Tai ruler, Mangrai (ruled *c.* 1259–1317; from 1292 to 1317 in Chiang Mai), conquered the ancient Mon kingdom of Haripunjaya and built a new capital at Chiang Mai. Under Mangrai and his successors, Lan Na—with Chiang Mai as its capital—became not only powerful but also a centre for the spread of Theravāda Buddhism to Tai peoples in what are now northeastern Myanmar, southern China, and northern Laos. Under Tilokaracha (ruled 1441–87), Lan Na became famous for its Buddhist scholarship and literature. During the 16th century Lan Na was conquered by the Myanmar and incorporated into the Burman empire. Subsequently, the central Tai (Siamese) states of Ayutthaya (Ayudhya, or Ayuthia) and Bangkok challenged Burman control over the area, but it was not until the 19th century that Lan Na was brought fully under Siamese rule.

The Ayutthayan period, 1351–1767. Whereas Sukhothai was an independent kingdom for only about 200 years, its successor, Ayutthaya—situated in the rich rice plains of the Chao Phraya River basin, about 55 miles

north of present-day Bangkok—lasted more than 400 years. During the Ayutthayan period the Tai consolidated their position as the leading power in what is now central and north-central Thailand, as well as throughout much of its southern peninsular region. Since many of Ayutthaya's neighbours called the country "Siam," or a name similar to it, the Tai of Ayutthaya came to be known as the Siamese.

Ayutthaya (Ayudhya) kingdom, mid-15th century.

Ayutthaya at first was only a small city-kingdom on the northwestern edge of the powerful Khmer empire. Within less than a century, however, Tai kings succeeded in pushing back the Khmer, and in 1431 they sacked their great capital of Angkor. Wars against neighbouring powers remained endemic, however, throughout the Ayutthayan period. In 1438 a greatly weakened Sukhothai was made a province of Ayutthaya. Lan Na, however, remained free of Ayutthayan control, although it was later brought under Burman influence.

The sack of Angkor

When the Siamese conquered Angkor, they brought many Khmer captives back to Ayutthaya with them. Some had been officials or craftsmen at the Khmer royal court. From them Ayutthaya's rulers adopted many Hindu practices that had been followed by the Khmer, including the concept of the ruler as god-king (*devarāja*). The king acquired powers of life and death over all his people. None but members of the royal family might gaze upon his face. He could be addressed only in a special language used exclusively for royalty, while those speaking to the king referred to themselves as "the dust beneath your majesty's feet."

The power of the ruler was enhanced not only through symbolic and ideological concepts drawn from Khmer-Hindu beliefs about the god-king but also through the centralization of political power. Trailok (ruled 1448–88) created a state in which the ruler stood at the centre of a series of concentric circles. As in the *müang* system, the outer circles were governed by hereditary lords, or *chao*. The inner circles, however, were administered by office-holders appointed by the king, and thus these operated to a limited degree on bureaucratic rather than hereditary lines.

The kings of Ayutthaya also issued formal codes of civil and criminal law based on the ancient Indian body of jurisprudence called the *Dharma-śāstra*. At the same time, a formal and highly complex hierarchical system assigned each person a varying number of units (called *sakdi na*) that designated one's rank within society. At the bottom of the scale, a slave was worth 5 units; freemen were ranked at 25 and above, while the heir apparent was assigned no fewer than 100,000 units.

The mass of the people in Ayutthayan times were peasant farmers, either freemen or slaves. The latter included war captives, bondsmen, and debtors. Freemen were obliged to work for six months each year for the local representatives of the king, to pay taxes, and to provide military service as required. An intricate patronage system extended throughout society, whereby clients provided their patrons with services in return for the protection of the patrons. Ayutthaya was an underpopulated society, and the constant need for manpower helped protect clients from excessive demands by patrons; if the demands of the patrons became too burdensome, the freeman could always move and take up new land as a last resort.

Despite the introduction of Brahmanism into court ritual and the admixture of animism and superstition that pervaded religious practice at all levels of society, Theravāda Buddhism took deep root throughout Siam during Ayutthayan times. The Buddhist monastic establishment played an important role in society, forming a focal point for village life, providing young males with an education, and offering those who elected to remain in the monkhood (sangha) a channel for upward social mobility.

Ayutthaya at its height was one of the wealthiest and most cosmopolitan cities of its day. Although it lay inland, it was easily accessible to oceangoing vessels traveling up the Chao Phraya River, and it became a thriving international trade emporium. It was during this period that European traders and travelers first started coming to Siam. The Portuguese reached Siam as early as 1511, following their conquest of Malacca (Melaka) on the Malay Peninsula; they were followed in the 17th century by Dutch, English, Spanish, and French traders and missionaries. Ayutthayan kings permitted settlements of Chinese, Indian, and Persian, as well as European, traders; they employed Japanese warriors and allowed Western missionaries to preach within Ayutthayan domains. In addition to engaging in extensive trade with China, Southeast Asia, and India, the rulers of Ayutthaya also sent triennial tribute missions to the Chinese imperial court, established Buddhist missions in Sri Lanka, and sent emissaries abroad as far afield as Europe. King Narai (ruled 1656–88) initiated a series of diplomatic exchanges between Ayutthaya and the French court at Versailles and even appointed a Greek adventurer, Constantine Phaulkon, as his chief minister. Eventually, however, the Europeans became overly zealous in their efforts to convert Buddhist Siamese to Christianity. In 1688 the Siamese expelled the French from Ayutthaya and all but closed their doors to the West for the next 150 years.

European contact

The primary threat to Ayutthayan sovereignty came not from Europe, however, but from Myanmar. In 1569 a force from the Burman state of Toungoo overran Ayutthaya and devastated the country for miles around. Led by Naresuan (ruled 1590–1605), Ayutthaya recovered its independence. Conflict with Myanmar persisted, however, and in the mid-18th century Burman armies once again captured Ayutthaya. This time the city was not to recover. Following the sacking of the city in 1767, the king and members of the royal family, along with thousands of captives, were deported to Myanmar. All Ayutthayan records were burned and its works of art destroyed.

The Thon Buri and Early Bangkok periods. A new era in Thai history began with the rise to power of Taksin, a military commander of great skill and charismatic personality who succeeded in pushing back the Burmans and seizing political power. In 1767 Taksin established his new capital at Thon Buri (Thonburi), on the opposite side of the Chao Phraya River from modern Bangkok. The new location was less accessible to the armies of Myanmar

than Ayutthaya had been and was ideally situated for the conduct of seaborne trade and commerce. Capitalizing on the trade relations that Siam had already developed with China, Taksin encouraged Chinese merchants and craftsmen to take advantage of the economic opportunities offered by the site of his new capital. Large numbers of Chinese settled permanently in Siam, where their involvement in business and trade—coupled with the tax revenues that these activities provided—helped restore the country's devastated economy.

Taksin not only recovered the territories that had formerly been part of the Ayutthayan empire but set out to extend Siamese control over new areas. His armies annexed part of what is now northeastern Cambodia and advanced up the Mekong River as far as present-day Vientiane, Laos. In the south they subdued the northern part of the Malay Peninsula, while to the north they pushed the Burmans out of the old northern Tai kingdom of Lan Na.

Within a few years of seizing power, however, Taksin showed signs of serious mental instability, and in 1782 he was overthrown. He was succeeded by his former military commander, known by his official name of Chao Phraya ("Great Lord") Chakri. As Rama I (ruled 1782–1809), he became the first king of the still-reigning Chakri (or Chakkri) dynasty.

Rama I

The early Chakri kings and a resurgent Siam. One of Rama I's first acts was to move his capital across the Chao Phraya River to Bangkok, which at the time was still a small village. By the mid-19th century, Bangkok had become a city of some 400,000 people, swelled by the huge numbers of Chinese who had poured into Siam during these years. In addition to settling in Bangkok, the Chinese established trading settlements inland, some of which grew into small towns. The Chinese thus gained control over both the internal and foreign trade of the country.

Myanmar continued to harass Siam throughout the early Chakri reigns. In 1785 it launched a massive invasion of the country, which was defeated only with great difficulty. Other lesser attacks followed. Not until the 1820s, when British encroachment on Myanmar forced Burman attention inward, was Siam able to relax its vigilance along its western borders. In the east Rama I and later Rama III (ruled 1824–51) reduced Khmer territories to vassal status, while in the south Rama III strengthened Siamese control over tributary states of the Malay Peninsula. Rama III also put down a major uprising in the north under Chao Anou, the young Lao ruler of the kingdom of Vien Chan (Vientiane). In 1827, Siamese armies razed and plundered Vientiane; thousands of Lao were taken prisoner and deported to central Siam.

The early Chakri kings sought to restore the cultural heritage of Ayutthaya. New temples and palaces were built following the same styles and even using the same bricks that had embellished Ayutthaya. Rama I reestablished court rituals, issued comprehensive law codes and authoritative Buddhist texts, and helped revive the sangha by placing learned and pious monks in leading positions within the Buddhist hierarchy. The early Bangkok period was also one of great literary flowering. The *Ramakian,* the Thai version of the Indian epic *Rāmāyaṇa,* was set to verse during the reign of Rama I, and the popular Chinese novel *San-kuo chih yen-i* (*Romance of the Three Kingdoms*) was translated into Thai. Rama II (ruled 1809–24), an accomplished poet, was a noted patron of the arts; Sunthon Phu, Thailand's greatest poet, wrote some of his best-known works during Rama II's reign.

Western influence also grew in mainland Southeast Asia during the early years of the 19th century, and with it came increasing Western pressures on Siam. When Britain declared war on Myanmar in 1824, Rama III feared that the British might also attack Siam. He subsequently agreed to sign the Burney Treaty (1826), which set conditions for the conduct of trade between the two countries.

Mongkut and the opening of Siam to the West. Demands for free trade and diplomatic representation in Siam accelerated with the British advances into Myanmar and Malaya and the opening of several Chinese ports following the first Opium War with China (1839–42). In 1855 Queen Victoria sent Sir John Bowring as her personal emissary to Siam to push for an end to all trade restrictions and to secure the rights to establish a British consulate in Bangkok and to set up separate law courts to try cases involving British subjects (extraterritoriality). The resulting Bowring Treaty (1855), in which Siam acceded to these demands, was followed shortly by similar treaties with other major European powers and with the United States. Although these treaties left Siam intact politically, they severely reduced the country's sovereignty and independence.

The Bowring Treaty

The opening of Siam to world trade and the development of a cash economy brought major changes to the country. The Bowring Treaty deprived the Siamese government of large sums in customs duties, one of its major traditional sources of revenue, forcing it to increase taxes in their stead. Large areas of the Chao Phraya basin were planted in rice and other cash crops for the world market, while the need to transport goods from the interior to the port of Bangkok led to the growth of canal systems and marketing networks.

The years following the Bowring Treaty were also marked by an increase in foreign influence in Siam. King Mongkut (Rama IV; ruled 1851–68) appointed several Western advisers and assistants to his court, including the Englishwoman Anna Leonowens, who became tutoress to his children. She later published her romanticized and inaccurate depiction of Mongkut's court. Foreign nationals began to take up long-term residence in Bangkok. Missionaries, although largely unsuccessful in converting Siamese to Christianity, set up the first Western medical facilities, secular schools, and printing presses in the country. Mongkut took great interest in the new Western ideas that were beginning to come into the country. He studied Latin, mathematics, and astronomy with the scholarly French Roman Catholic missionary Jean-Baptiste Pallegoix and English with American Protestant missionaries, one of whom, Dan Beach Bradley, later founded the country's first newspaper.

Mongkut was already 46 years old when he came to the throne. He had spent 26 years in the monkhood, during which time he had become a keen scholar of Pāli (the language of the Theravāda texts) and an expert in Buddhist doctrine. Mongkut also had become concerned that many superstitious practices had grown up around the core Theravāda teachings, and he established the reformed Thammayut sect, which was based on the purification of Buddhist practice. The Thammayut order later came to dominate the Thai monkhood. Although Mongkut was an absolute monarch, he began to break down the age-old tradition of treating the king as a god. He traveled widely throughout his kingdom, inquiring about the conditions of his subjects. He also was the first Siamese monarch to allow his subjects to gaze directly upon his face. Mongkut's willingness to adapt traditional Siamese patterns to make way for more modern ideas helped pave the way for the more profound social and political changes that were to take place in Siam under his successor.

Chulalongkorn and the foundations of modern Thailand. Mongkut was succeeded by his 15-year-old son Chulalongkorn (Rama V; ruled 1868–1910). Because of Chulalongkorn's youth, the country was ruled by a regent until the prince came of age in 1873. Chulalongkorn was faced with continuing Western pressure, and he maintained his father's policy of making territorial concessions to the West in the hope of retaining Siam's overall independence. In 1893, after French gunboats forced their way up the Chao Phraya River to Bangkok, he was forced to cede all Lao territories east of the Mekong River to France, and in 1907 the French took over three territories in northwestern Cambodia and Lao territory west of the Mekong that had been under Siamese suzerainty. Two years later the Siamese government lost rights over four Malay states to the British.

Cession of territory to France

At the same time as he sought to fend off the Western powers from without, Chulalongkorn undertook major reforms within the country. These were often difficult to achieve, since they undercut the power bases of influential men at court. The young king proceeded gradually, assisted by several of his brothers and half-brothers; many

of these—in particular the brilliant and energetic Prince Damrong Rajanubhab—were men of outstanding ability. The internal reforms carried through during Chulalongkorn's reign included the reorganization of the government into ministries with functional responsibilities and the creation of a centralized bureaucracy, the institution of a uniform and centralized system of administration over the outlying provinces, the systematization of government revenue collection, the abolition of slavery and labour-service requirements, the establishment of law courts and reformation of the judiciary, the introduction of a modern school system, and the construction of railways and telegraph systems. In addition, he backed a major reorganization of the Buddhist monkhood, bringing all monks throughout the country into the sangha as a nationwide religious hierarchy that was linked at its apex to the king. By any standards, the sheer scale of Chulalongkorn's reforms are remarkable, and his reign is commonly regarded as one of the greatest in Thai history.

Vajiravudh and Prajadhipok: the last absolute monarchs of Siam. Chulalongkorn's policies were continued by his sons Vajiravudh (Rama VI; ruled 1910–25) and Prajadhipok (Rama VII; 1925–35). In 1917 Vajiravudh, the first Thai monarch to be educated abroad, opened Thailand's first university, which he named for his father. In 1921 he made universal primary education compulsory throughout the nation. He also passed an act aimed at assimilating the growing number of Chinese entering the country, which declared that all students must be taught to read, write, and understand standard Thai (Siamese) and be instructed in the duties of being a good Siamese citizen. Vajiravudh is noted principally, however, for promoting Thai nationalism. In his voluminous writings he constantly stressed the need for his subjects to be loyal to nation, religion, and king. He not only strengthened the army and navy but created a paramilitary organization, the Wild Tiger Corps, that was independent of the regular army. In 1917 he took Siam into World War I on the side of the Allies, and after the war he succeeded in persuading the Western powers to give up their extraterritorial rights in Siam. Vajiravudh also passed a law in 1913 that required all Siamese to adopt surnames, and he encouraged his people to take on more modern clothing styles and to abandon such habits as chewing betel nuts.

Vajiravudh also was notorious for his extravagance, and his successor, Prajadhipok, inherited serious fiscal problems from his brother. The new king ordered layoffs throughout most government departments, both at the start of his reign and again during the Great Depression of the 1930s. The cuts caused severe economic hardships to many government officials and their families and were among the many reasons for the rise of popular discontent with the monarchy during his reign. In addition, a growing middle class was becoming unhappy with the domination of the government by members of the royal family and with the absence of wider participation in political decision making, and a popular press emerged that was able to give voice to these discontents.

Rise of popular discontent

The 1932 coup and the creation of a constitutional order. One focus of civilian discontent centred around a group of students who were educated overseas and were deeply dissatisfied with the tight political control that Siam's ruling families held over the country. Some of these students became politically radicalized during the course of their education in Europe in the 1920s and early 1930s. They were led by Pridi Phanomyong, a brilliant young lawyer studying in Paris, who became the leader of an association of overseas Siamese students. He was closely associated with a career artillery officer, Luang Phibun Songkhram (Phibunsongkhram), who was then studying military science in France. In 1927 Pridi and Phibun formed the People's Party, which became the nucleus of a revolutionary group plotting to overthrow Siam's absolute monarchy. On their return to Siam the two men and their associates, who became known as the Promoters, built up a revolutionary following among students, low-level government officials, and military officers.

On June 24, 1932, while Prajadhipok was away from Bangkok, the Promoters staged a bloodless coup; they seized control of the army, imprisoned the royal officials who had constituted the ruling group, and persuaded the king to agree to rule under a constitution. Under the new government, a State Council and National Assembly were established. Many members of the new government had not played a direct part in the coup, and some were quite conservative in their political thinking. In early 1933, when Pridi drew up an economic plan for the country that was far more radical than many members of the new government could accept, feelings ran so high that the king was forced to suspend the National Assembly. Fearing that the royalists would regain control of the government, the military leaders forced the reconstitution of the Assembly. This was followed by an attempted royalist countercoup in October 1933 under Prince Boworadet (Bavoradej), a cousin of the king. Although there was no evidence of royal collusion, Prajadhipok found his position untenable. In early 1934 he left for England, and in March 1935 he abdicated. A regency council was appointed to act for his successor, Prince Ananda Mahidol, then a schoolboy studying in Switzerland, until he came of age.

Between 1933 and the end of 1938 the military grew ever stronger. The years just before World War II were marked by a tripling of the military budget, the establishment (1934) and subsequent spread of a paramilitary youth movement with fascist overtones, and a growing alliance with Japan.

The Phibun dictatorship and World War II. In December 1938 Phibun Songkhram took over as military dictator, and the following year he changed the name of the country from Siam to Thailand. He embarked on a strongly nationalistic policy that was chauvinistic and anti-Chinese at home and irredentist and pro-Japanese abroad. In November 1940, taking advantage of the defeat of France by Germany the previous June, Phibun invaded French territories in western Laos and northwestern Cambodia that formerly had been under Thai control. Japan supported Thai claims to the disputed lands.

Invasion of French territories

Thailand's leaders nonetheless sought help from Britain and France against an increasingly aggressive Japan, but the British were too deeply involved in Europe to provide them with meaningful support. On Dec. 8, 1941, just after the Japanese attack on Pearl Harbor, Japanese troops entered Thailand and requested the right of passage through the country to facilitate their planned surprise rear attack on British-held Singapore. After a brief fight against the advancing Japanese, all Thai troops were ordered by Phibun to lay down their arms, and Thailand signed a full Treaty of Alliance with Japan; in January 1942 the Thai government declared war on Britain and the United States.

Thailand gained minor territorial concessions in Burma and Malaya, as well as in Laos and Cambodia, from its wartime alliance with Japan, but the Thai economy suffered greatly, ultimately undermining public confidence in Phibun. From 1942 onward, overseas resistance groups based in the United States and Britain made contact with similar groups within Thailand led by Pridi Phanomyong. Collectively called the Free Thai, they conducted raids against the Japanese and succeeded in infiltrating the government. In July 1944 Phibun was forced to resign, and in August 1945 Japan surrendered.

The postwar crisis and the return of Phibun. The chief problem facing Thailand in 1945 was how to restore its international reputation in view of its wartime alliance with Japan. The United States, however, had never accepted Thailand's declaration of war, which it believed to have been signed under duress. Once Thailand returned the territories seized from France in 1940–41, it was admitted to the United Nations, and its standing in the international community was restored. The immediate postwar years, however, were not easy for Thailand. Phibun narrowly escaped trial as a war criminal and temporarily retired from public life. Then, in June 1946, the recently enthroned Ananda Mahidol was found dead of a gunshot wound, an event that shocked the nation. Pridi Phanomyong was accused—falsely—of regicide and felt obliged to go into exile. Political life degenerated into factionalism and disorder, and in April 1948 the military, again led by Phibun, seized power.

As the international political climate became one of Cold War, the West began to look to Thailand as a potential bastion against the rise of communism and the growing influence of the Soviet Union and China in Southeast Asia. Thailand sent troops to join the United Nations forces during the Korean War, and in 1954 it became a charter member of the Southeast Asia Treaty Organization, a regional anticommunist defense organization to which the United States pledged its support. The establishment of a communist regime in China in 1949 aroused Phibun's fears about the spread of communism within Thailand and led him to carry out a series of measures directed against members of the Chinese community. He also imprisoned many leaders from other groups whom he feared might try to secede from the Thai nation, in particular the Lao of northeastern Thailand and the Malays in the south.

Between 1951 and 1957 the United States poured large amounts of economic and military aid into Thailand to build the nation's infrastructure and boost its military and police forces. This massive financial support laid the basis for an economic boom in Thailand that has continued to the present day. Access to these funds also rendered the military largely independent of the political process; an alliance of convenience developed between the military rulers—headed by Phibun and the newly emerging army chief, Sarit Thanarat—and the police, in which the latter suppressed Phibun's political opponents in return for a share of the political spoils.

Economic boom

Sarit was entrusted by Phibun with the buildup and modernization of the Thai army, and by 1954 he had risen to the rank of field marshal. Like a number of upper-echelon military officials during this period, Sarit had become heavily involved in business activities and served on numerous corporate boards. Under the Phibun government, most of the country's small number of manufacturing firms were government-owned, while imports and exports were tightly controlled. Sarit and many members of the middle class, particularly businessmen of Chinese descent, were disappointed by the poor economic results that Phibun's policy of economic nationalism generated. Over the next three years, public confidence in the Phibun regime waned, and in September 1957 Sarit took over the government.

Military dictatorship, economic growth, and the reemergence of the monarchy. Sarit was dictator from 1958 until his death in 1963, during which time he instituted new economic policies that favoured both domestic and foreign private investment. His commitment to economic development, coupled with a massive rise in foreign economic and military aid to Thailand (especially from the United States), led to a marked rise in Thailand's gross national product. Not only were large amounts of money funneled into the military, but there was a major increase in the construction of highways, irrigation projects, electrification schemes, and schools. Sarit also encouraged Bhumibol Adulyadej, who had succeeded his brother as king in 1946, to make the public more aware of the monarchy. The king and queen made trips around the country and sponsored public service activities, and by 1960 they had become widely popular. The monarchy, which had been in eclipse since 1932, once again became a significant institution in Thailand.

Sarit was admired by many as a strong and decisive ruler, but his popularity diminished significantly after his death, when the extent of his personal corruption became widely known. The aura of corruption haunted his successors, Thanom Kittikachorn and Praphas Charusathian, who jointly held power throughout the decade following Sarit's death. Their term of office also was noted for the continuing growth of the Thai economy and for Thailand's increasing involvement with the United States during the Vietnam War. By 1969 Thailand had more than 11,000 troops serving in Vietnam. Huge sums of American money continued to pour into Thailand throughout the Thanom-Praphas years, further driving up the level of economic development but also contributing substantially to the growth of corruption and to a rising gap in the standard of living between rich and poor. Popular disaffection grew—particularly in the impoverished northeast and among alienated groups such as the Muslim Malays in the south and the Hmong in the far north—gradually crystallizing into outright insurgency.

The Vietnam War

The 1973 revolution and its aftermath. Faced with growing internal dissent, Thanom made halfhearted attempts to introduce minor democratic reforms before reimposing direct military rule in 1971. For many Thai, especially the growing number of middle-class citizens educated abroad and exposed to Western democratic ideas, this undermined their vision of the country's future. Students, in particular, felt betrayed by the failure of the government to provide for reform and held huge public demonstrations calling for the promulgation of a constitution. Violence between police and students escalated, and in October 1973 Thanom and Praphas were forced to call on the king to restore peace to the country. Following a royal plea, the students agreed to disperse, but on October 14 Thanom and Praphas were forced to leave the country.

For the first time since 1932, the monarchy assumed a direct role in Thai politics. The king chose Sanya Dharmasakti, a former rector of Thammasat University, to be interim prime minister and to oversee the drafting of a new constitution. The constitution, promulgated in 1974, ushered in a brief period of parliamentary democracy in Thailand. The open debates in parliament about policy issues, however, were interpreted as an indication of political instability by ranking members of the military, while the triumph of communist governments in Vietnam, Cambodia (Kampuchea), and Laos in 1975 were pointed to as a threat requiring a stronger Thai government. In October 1976 the military, this time with the backing of the king, once again took control of the government and abolished both parliament and the constitution.

The new coup polarized the Thai political system. Many of the students who had led or supported the movement of the early 1970s went into the jungle to join what previously had been a small, rural-based communist insurgency. By mid-1977, the Communist Party of Thailand was beginning to mount an increasingly effective challenge to the military-backed government. Fearing increasing unrest, the military leaders—in yet another October coup—ousted the extreme right-wing government they had installed a year before and handed over power to General Kriangsak Chomanand, who was open to a more democratic type of government.

Growing insurgency

The search for a new political order. By 1980, when General Prem Tinsulanonda replaced Kriangsak, Thailand had established a system of government in which the military and the parliament shared power. Prem, who served as prime minister from 1980 to 1988, eliminated the threat of the Communist Party of Thailand by declaring a general amnesty. Thailand, however, faced a new threat along its eastern border from Vietnam following its occupation of Cambodia in 1979. In 1988, Chatichai Choonhavan, leader of the Chat Thai political group, replaced Prem as prime minister. For the first time since 1976, Thailand had a government headed by an elected, rather than a military, leader, but the military retained a veto over major policy decisions.

Political instability continued to trouble the country in the early 1990s. At the beginning of 1991, Chatichai's government, already criticized for corruption, went too far in challenging the military and was toppled by a junta. Led by General Sunthorn Kongsompong and including army chief Suchinda Kraprayoon, the junta promised free elections (subsequently held in March 1992) and appointed as prime minister Anand Punyarachun, a liberal, former diplomat, and business leader. Anand sought unsuccessfully to remain independent of the military, and Suchinda succeeded him as prime minister in April 1992. Suchinda's appointment was strongly opposed by advocates of democracy. In May the army met the escalating antigovernment demonstrations with bloody repression; the king intervened to defuse the situation, after which Suchinda resigned. Anand headed a caretaker government until new elections could be held in September. Parties opposed to military rule won a majority of seats and chose as prime minister Chuan Leekpai, who became the longest-serving elected prime minister in Thai history. Since 1992 govern-

ments have been based upon parliamentary majorities. The new constitution of 1997, which recognized broader rights for citizens than any previous constitution, reflected the growing influence of nongovernmental organizations and the emergence of a new civil society in Thailand.

Economic and foreign-policy developments. From the 1960s to the 1990s Thailand had one of the world's fastest-growing economies. By the 1990s it was considered to be part of a second wave of so-called newly industrializing countries, or NICs (including other countries of the Pacific Rim such as Malaysia and Indonesia), following in the path of such first-wave NICs as South Korea and Taiwan. In July 1997, however, a financial crisis hit Thailand and swept across most of East and Southeast Asia. In Thailand, financial institutions collapsed or were taken over and the economy slowed to a near halt. Foremost among those responsible for the crisis were the many finance companies that made short-term loans to fund long-term property investments.

From the 1950s through the '70s, Thailand's foreign policy was based on anticommunism and a special relationship with the United States. The withdrawal of U.S. forces from Vietnam, the communist takeovers in Vietnam, Cambodia, and Laos, the disintegration of the Soviet Union, and the end of the Cold War spurred Thailand to reassess its foreign policy. Since the 1980s the focus has shifted from security to trade, ties with the United States have weakened while those with Japan and China have been strengthened, and Thailand has encouraged regional economic relationships, even with its former enemies—Vietnam, Laos, and Cambodia. (E.J.K./C.F.Ke.)

Vietnam

The Socialist Republic of Vietnam (Vietnamese: Cong Hoa Xa Hoi Chu Nghia Viet Nam) is a densely populated nation occupying the eastern part of the Indochinese Peninsula. It has an area of 127,800 square miles (331,000 square kilometres). From north to south it extends about 1,025 miles (1,650 kilometres) and at its narrowest part is about 30 miles wide. Vietnam is bordered by China to the north, the South China Sea to the east and south, the Gulf of Thailand to the southwest, and Cambodia and Laos to the west. The capital, Hanoi, is located in the north, while the country's largest city, Ho Chi Minh City (formerly Saigon), is in the south. The current Vietnamese nation was established in July 1976, after a period of prolonged internal warfare that was fueled by Cold War politics and after being partitioned (1954–75) first militarily and later politically into the Democratic Republic of Vietnam, better known as North Vietnam, and the Republic of Vietnam, usually called South Vietnam.

PHYSICAL AND HUMAN GEOGRAPHY

The Annamese Cordillera

The land. *Relief.* Vietnam's principal physiographic features are the Annamese Cordillera (French: Chaîne Annamitique; Vietnamese: Nui Truong Son), extending from north to south in central Vietnam and dominating the interior, and two extensive alluvial deltas formed by the Red (Hong) River in the north and the Mekong (Cuu Long) River in the south. Between these two deltas is a long, relatively narrow coastal plain.

From north to south the uplands of northern Vietnam can be divided into two distinct regions—the region north of the Red River and the massif that extends south of the Red River into neighbouring Laos. The Red River forms a deep, relatively wide valley that runs in a straight northwest-southeast direction for much of its course from the Chinese border to the edge of its delta. North of the Red River the relief is moderate, with the highest elevations occurring between the Red and Lo (Clear) rivers; there is a marked depression from Cao Bang to the sea. In the Red River delta and in the valleys of the region's other major rivers are found wide limestone terraces, extensive alluvial plains, and low hills. The northeast coast is dotted with hundreds of islands composed mostly of limestone.

Compared with the area north of the Red River, the vast massif extending southwest across Laos to the Mekong River is of considerably higher elevation. Among its outstanding topographic features is Fan Si Peak, which at 10,312 feet (3,143 metres) is the highest point of elevation in Vietnam. South of the Black (Da) River are the Ta P'ing, Son La, and Moc Chau plateaus, which are separated by numerous deep valleys.

In central Vietnam the Annamese Cordillera runs parallel to the coast, with several peaks rising to elevations of more than 6,000 feet. Several spurs jut into the South China Sea, forming sections of the coast isolated from one another. Communication across the central ranges is difficult. The southern portion of the Annamese Cordillera has two identifiable regions. One consists of plateaus of approximately 1,700 feet in elevation that have experienced little erosion, as in the Dac Lac Plateau near Buon Me Thuot. The second region is characterized by heavily eroded plateaus: in the vicinity of Pleiku, the Kontum Plateau is about 2,500 feet above sea level, and, in the Da Lat area, the Di Linh Plateau is about 4,900 feet.

Drainage. Below the northern uplands is the Red River delta. Roughly triangular in shape, it extends some 150 miles inland and measures 75 miles along the Gulf of Tonkin. The delta can be divided into four subregions. The northwestern section has the highest and most broken terrain, and its extensive natural levees invite settlement despite frequent flooding. The low-lying eastern portion has benchmarks of less than seven feet above sea level in the vicinity of Bac Ninh. Rivers there form small valleys only slightly lower than the general surface level, and they are subject to flooding by the area's unusually high tides. The third and fourth subregions consist, respectively, of the poorly drained lowlands in the west and the coastal area, which is marked by the remains of former beach ridges left by the continuous expansion of the delta.

The Annamese Cordillera forms a drainage divide, with rivers to the east flowing to the South China Sea and those to the west to the Mekong River. South of the mountain range there is an identifiable terrace region that gives way to the Mekong River delta. The terrace region includes the alluvial plains along the Saigon and Dong Nai rivers. The lowlands of southern Vietnam are dominated by alluvial plains, the most extensive of which is the Mekong River delta, covering an area of 15,400 square miles in Vietnam. Smaller deltaic plains also occur along the south-central coast of the South China Sea.

The Mekong River delta

Soils. In northern Vietnam the heavy monsoonal rains wash away rich humus from the highlands, leaving slow-dissolving alumina and iron oxides that give the soil its characteristic reddish colour. The soils of the Red River delta vary: some are fertile and suitable to intense cultivation, while others lack soluble bases that make the growing of crops feasible. Nonetheless, the delta soils are easily worked. The diking of the Red River to prevent flooding deprives the delta's rice fields of enriching silts, necessitating the use of chemical fertilizers.

There are some two dozen soil associations, but certain soil types predominate. Among these are red and yellow podzolic soils (*i.e.*, soils that are heavily leached in their upper layers, with a resulting accumulation of materials in the lower layers), which occupy nearly half of the land area, and lateritic soils (reddish brown, leached tropical soils), which constitute about 10 percent. These soil types dominate the central highlands. Alluvial soils account for about one-fourth of the land in the south and are concentrated in the Mekong River delta, as are peat and muck soils. Gray podzolic soils are found in parts of the central highlands and in old terraces along the Mekong River, while regurs (rich black loams) and lateritic soils occur in both the central highlands and the terrace zone. Along the coast of central Vietnam are regosols (soft, undeveloped soils) and noncalcic brown soils.

Climate. The northern part of Vietnam is on the edge of the tropical climatic zone. During January, the coldes month of the year, Hanoi has a mean temperature of 63° F (17° C), while the annual average temperature is 74° F (23° C). Farther south, the average annual temperature in Hue is 77° F (25° C) and in Ho Chi Minh City is 81° F (27° C); in the highland city of Da Lat, it drops to 70° F 21° C). The winter season in northern Vietnam lasts

from November to April; from early February to the end of March there is a persistent drizzling rain, and March and April sometimes are considered to be a transitional period. The summer in northern Vietnam lasts from April or May to October and is characterized by heat, heavy rain, and occasional typhoons. In central and southern Vietnam the southwest monsoon winds between June and November bring rains and occasional typhoons to the eastern slopes of the mountains and the lowland plains. Between December and April there is a drier period that is characterized by winds of the northeast monsoon and, in the south, by high temperatures.

Plant life. The vegetation of Vietnam is rich and diversified, reflecting the great range of climate, topography, and soils and the varying effects of human habitation.

Forest categories

The forests of Vietnam can be divided into two broad categories: evergreen forests, which include conifers, and deciduous forests. There are more than 1,500 species of woody plants in Vietnam, ranging from hardwoods such as ebony and teak to palms, mangroves, and bamboos. There also are numerous species of woody vines (lianas) and herbaceous plants. In the aggregate, the dense and open forests, savannas, brushland, and bamboo cover approximately half of the total area of Vietnam.

In most areas the forests are mixed, containing a great variety of species within a given area. Rain forests are relatively limited, and pure stands are few. The nearest to pure forest types are the pines—the three-needled *Pinus khasya* and the two-needled *P. merkusii* found in the uplands—and the mangrove forests of the coastal areas. In the mountainous regions are subtropical species from such genuses as *Quercus, Castanopsis, Pinus,* and *Podocarpus.* Brushwood, bamboo, weeds, and tall grasses invade logged areas and grow around settlements and along arterial highways and railroads. Between the logged areas and the upland forests are other mixtures of forest types.

A large part of the forest in the central highlands is dense and rich in broad-leaved evergreens and semievergreens, some of which produce valuable timbers. Some of this region still is composed of undisturbed (primary) forests. Other types of forests there include secondary forests; open forests, which typically have trees of the Dipterocarpaceae family and species from the genus *Lagerstroemia* (crape myrtle); mangrove forests; and barren lands of sand dunes with eucalyptus and small, thorny deciduous trees and species from the *Casuarina* genus of flowering plants. Cogon grass (*Imperata cylindrica*) is commonly found in the open forests, and savannas occupy large areas formerly covered by forests. Grass and sedge swamps are characteristic of the Thap Muoi Plain (Plain of Reeds), a depression in the Mekong River delta.

During the Vietnam War, herbicides were used by the U.S. Army to defoliate large areas of forest in southern Vietnam. Most of these forests have been regenerating, however, and resettlement programs and illegal logging appear to have created longer-lasting damage.

Animal life. The most common domesticated animals in Vietnam are water buffalo, cattle, dogs, cats, pigs, goats, ducks, and chickens. Wild game in the central highlands includes elephants and tapirs; rhinoceroses once roamed there, but none have been seen since the early 1940s. Also found in the forests are large cats, including tigers, leopards, and ounces (snow leopards); several kinds of wild oxen, including gaurs and koupreys; and various types of bears, among them black bears and sun bears (honey bears). Deer are plentiful and include the small musk deer and barking deer. Other common wild animals are wild boars, porcupines, jackals, otters, mongooses, hares, skunks, and squirrels, including flying squirrels.

There are many different kinds of small wildcats and three types of civets—Malagasy civets, binturongs, and palm civets. Primates such as the langur, macaque, gibbon, and rhesus monkey live in the forests. Crocodiles are found on the edges of some lakes and along riverbanks; other reptiles include several kinds of lizards, pythons, and cobras. Of the wide variety of land and water birds, some 600 species have been identified in southern Vietnam alone.

Traditional regions. Diverse cultural traditions, geographic variations, and historical events have created distinct traditional regions within the country. The general topographic dichotomy of highland and lowland regions also has ethnolinguistic significance: the lowlands generally have been occupied by ethnic Vietnamese, while the highlands have been the home of numerous smaller ethnic groups that differ culturally and linguistically from the Vietnamese. The highland peoples can be divided into the northern ethnic groups, with affinities to peoples in southern China, and the southern highland populations, with ties to the Mon-Khmer and Austronesian peoples of Cambodia, Indonesia, and elsewhere in Southeast Asia. A north-south variation also evolved among the ethnic Vietnamese as they expanded southward from the Red River delta along the coastal plain and into the Mekong River delta. After the mid-19th century, Vietnam was divided by the French into Tonkin in the north, Annam in the centre, and Cochinchina in the south. The Vietnamese themselves have long made a distinction between the northern region, with Hanoi as its cultural centre; the central region, with the traditional royal capital of Hue; and the southern region, with Saigon (Ho Chi Minh City) as its urban centre.

Highland and lowland dichotomy

A settlement (background) overlooking rice paddies near Hoa Binh, northern Vietnam.

Settlement patterns. There are several distinct rural settlement patterns in Vietnam. Especially in northern and central Vietnam, geomantic principles influenced the traditional internal orientation of houses and community buildings; in central Vietnam, buildings often faced the sea. In the densely populated Red River delta, villages often are tightly nucleated settlements, usually enclosed by a bamboo hedge or an earthen wall. Those along rivers, canals, or roads often abut each other, forming a single elongated settlement. Lowland Vietnamese villages on the central coastal plain are characteristically close-knit, small clusters of farmsteads near watercourses. Fishing villages often are situated in sheltered inlets. In the Mekong delta many settlements are strung out along waterways and roads; most are loose-knit clusters of farmsteads, with some farmsteads scattered about the rice fields. The pattern of settlements of the Cham and Khmer minorities closely resembles that of the Vietnamese. Most highland peoples build their houses on pilings.

Vietnam's traditional major cities are Hanoi, Hue, and Saigon (Ho Chi Minh City). Throughout Vietnamese his-

tory the Hanoi area has been important and was the site of several early capitals. Hanoi also served as the French capital of Indochina from 1902 until 1954, and the city retains the architecture of that heritage. The city's port of Haiphong was developed by the French in the late 19th century as a trade and banking centre. Hue was the seat of the Nguyen family, which controlled central and southern Vietnam from the late 17th to the late 19th century. Located on the Huong (Perfume) River, it was laid out in the early 19th century as a political and religious centre, and its economic functions were ancillary. Saigon was built largely by the French in the second half of the 19th century as the administrative capital and principal port of Cochinchina. The city's architecture recalls towns and cities in southern France. The adjoining city of Cholon long has been a major centre for ethnic Chinese.

The people. *Ethnolinguistic groups.* Vietnam has one of the most complex ethnolinguistic patterns in Asia. The Vietnamese were significantly Sinicized during a millennium of Chinese rule. Vietnamese, one of the Mon-Khmer languages of the Austro-Asiatic language family, exhibits strong Chinese influence.

The Cham

Indian influence is found among the Cham and Khmer minorities. The Cham, whose language belongs to the Austronesian language family, formed the majority population in the Indianized kingdom of Champa in what is now central Vietnam from the 2nd century to the late 15th century AD. Small numbers of Cham remain in the south-central coastal plain and in the Mekong delta near the Cambodian border. The Khmer (Cambodians), whose language is one of the Mon-Khmer languages, are scattered throughout the Mekong delta.

Many other ethnic groups inhabit the highlands. While cultures vary considerably in the central highlands, shared characteristics include a traditional way of life still largely oriented around kin groups and small communities. Known collectively by the French as Montagnards ("Highlanders"), these peoples have affinities with other Southeast Asians. Many groups—such as the Rade (Rhade), Jarai, Chru, and Roglai—speak Austronesian languages, linking them to the Cham, Malay, and Indonesian peoples; others—including the Bru, Pacoh, Katu, Cua, Hre, Rengao, Sedang, Bahnar, Mnong, Mang (Maa), and Stieng—speak Mon-Khmer languages, affiliating them with the Khmer. Highlanders have experienced little Chinese or Indian influence, but they were exposed to Western (French and then American) influence from the late 19th century until the early 1970s. French missionaries and administrators provided roman script for some of the Montagnard languages, and additional orthographies have been devised since. The Montagnards have exhibited an intense desire to preserve their own cultural identities.

The various groups in the uplands of northern Vietnam have ethnolinguistic affiliations with peoples in Thailand, Laos, and southern China. The largest of these are the tribal Tai (Thai) groups who speak Tai languages and generally live in upland valleys. Hmong (Miao, or Meo) and Mien groups, who speak languages of the Sino-Tibetan language family, are scattered at higher elevations.

Religions. Confucianism, Taoism, and Mahāyāna Buddhism flowed into Vietnam over many centuries. Gradually they became intertwined, simplified, and Vietnamized to constitute, along with vestiges of earlier animistic beliefs, a Vietnamese folk religion that came to be shared to some considerable extent by all Vietnamese, regardless of region or social class. Animistic beliefs are held by many tribal peoples. During the 1920s the syncretic religion of Cao Dai appeared, and in the 1930s the Hoa Hao neo-Buddhist sect spread through parts of the Mekong delta.

Roman Catholicism was introduced into Vietnam in the 16th century and spread rapidly following the French conquest in the mid-19th century. The heaviest concentrations of Roman Catholics in Vietnam once were in the north, but many fled to the south after the partition of the country in 1954. Protestantism came to Vietnam in 1911 and spread mainly among small segments of the urban population in the central and southern regions.

In 1954 all foreign Roman Catholic clergy were expelled from North Vietnam, leaving only native priests. The North Vietnamese government tried to supplant organized religion with its own patriotic Buddhist, Cao Dai, Catholic, and Protestant religious organizations; Catholic clergy and membership renounced their allegiance to Rome. With the conquest of South Vietnam by North Vietnam, all foreign Christian clergy were expelled. The country's current constitution has guaranteed freedom of religion, though in practice government controls have been relaxed only gradually.

Demographic trends. Vietnam's population has grown rapidly since reunification in 1975. As a result, an increasing proportion of the population is young.

Migration pattern

The migration pattern long has been predominantly from north to south, and more recently there also has been migration from the lowlands to higher elevations and from rural to urban areas. In 1954 nearly one million people moved from north to south. In both the north and the south in the late 1950s, there were programs to resettle ethnic Vietnamese from the lowlands to the uplands. While these programs were discontinued in the south in 1963, they continued in the north; between 1976 and 1980 they were revived throughout the country and greatly intensified, with a significant number of people moving from the south to the central highlands. Since then, however, there has been an overall flow of migrants into Ho Chi Minh City and its environs, as well as into the central highlands. Out-migration has been greatest in parts of the northeast and along the central coastal plain.

Emigration also has been considerable since reunification. Between 1975 and 1990 hundreds of thousands of Vietnamese left the country, both legally and illegally, and an unknown number died at sea. Many have remained in refugee camps in Thailand and other countries, but a large number have emigrated, especially to the United States.

The economy. Vietnam's greatest economic resource is its literate and energetic population. Its long coastline provides excellent harbours, access to marine resources, and many attractive beaches and areas of scenic beauty that are well suited to the development of tourism; a lack of infrastructure, however, has inhibited full utilization of these assets. The actual potential for economic growth based on Vietnam's wealth of natural resources, however, is being rendered increasingly problematic by population growth, environmental degradation, and rising domestic demand, and the country remains one of the poorest in the world.

During the period of 1954–75, when the country was divided, there were three layers to the Vietnamese economy: a bottom layer based on the cultivation of rice, a middle layer dominated by mining in the north and rubber plantations in the south, and a third layer that was a wartime creation marked by large-scale Soviet and Chinese aid in the north and substantial American aid in the south. In the north, land reform in 1955–56 was followed by rapid collectivization of agriculture and handicrafts. Government investment favoured heavy industry at the expense of agriculture, handicrafts, and light industry, the traditional mainstays of the economy. Heavy industry grew, but efficiency was low, quality was poor, and further progress was hampered by deficiencies in agriculture and light industry. Economic aid from socialist countries masked many economic deficiencies. The southern economy was largely based on free enterprise, with significant state ownership of industrial enterprises. Agriculture flourished in the Mekong delta, while trade and transport were developed by private enterprise. The standard of living was significantly higher in the south than it was in the north.

After reunification, the northern model of development was imposed on the entire country. Efforts to socialize the commercial sector and to collectivize agriculture met with resistance, especially in urban centres and in the rich Mekong delta, where the majority of farmers in the 1970s were self-sufficient, middle-income peasants. The south also experienced a severe loss of human resources. Many well-educated people fled Vietnam after 1975. Hundreds of thousands more, mainly those associated with the former government or the Americans, were placed in jails or reeducation centres, while other skilled but politically suspect people were forced to resettle in remote areas. Efforts to abolish private enterprise in the south and deteriorating

Population loss in the south

political relations with China mainly affected the ethnic Chinese and precipitated a flight of ethnic Chinese from Vietnam in 1978. Large police and military expenditures further strained the budget and diverted resources from productive enterprises.

These factors, combined with poor management of state-run programs, precipitated a severe economic crisis. Food production and per capita income dropped, and consumer goods were shoddy, expensive, and in short supply. The government responded with minor reforms in 1979 and more basic changes beginning in 1986. Vietnam began to move away from a state-controlled, centrally planned, subsidized economy toward one that utilized market forces and incentives and tolerated private enterprise—albeit under continuing government control. In response, the quality and variety of food and of various consumer goods increased, as did exports.

Resources. Mineral deposits, mainly in the north, are diverse. There are large reserves of anthracite coal, as well as of phosphates, high-grade chromite, tin, antimony, bauxite, gold, iron ore, lead, tungsten, zinc, and lime. A number of offshore oil deposits have been discovered in the South China Sea, mainly off Vietnam's southern coast.

Agriculture, forestry, and fishing. Agriculture is by far the most important economic sector in Vietnam. The great majority of the population earns its income from farming. In addition, agriculture is the main source of raw materials for the processing industries and a major contributor to exports; by the late 1980s Vietnam was again exporting rice after years of shortages. Permanent cultivation covers large areas of the country's lowlands and smaller portions of the highlands. The primary agricultural areas are the Red River delta, the Mekong River delta, and the southern terrace region. The central coastal land, which is subject to destructive typhoons, is a region of low productivity. The central highlands area, traditionally one of low productivity, has been intensively cultivated since 1975, but with mixed results.

Rice is the most important crop. It is grown on about four-fifths of the cropped land, principally in the Red River and Mekong River deltas. Other major food crops are cassava (manioc), corn (maize), soybeans, peanuts (groundnuts), and sweet potatoes. Agriculture is highly labour-intensive in Vietnam, and much plowing is still done by water buffalo. There are many plantations of banana and other fruit trees, coconut, and sugarcane, most of them found in the Mekong River delta and the southern terrace regions. Coffee and tea are grown in the central highlands. The production of rubber was disrupted by the war but has been restored in the central highlands and southern terrace regions. Other cash crops include tobacco and jute. Fields, groves, and kitchen gardens throughout Vietnam include a wide variety of fruit trees (banana, orange, mango, jackfruit, and coconut) and vegetables. Kapok trees are found in many villages, and the Vietnamese cultivate areca palms and betel peppers for their nuts and leaves and mulberry bushes to feed silkworms.

Seafood export

The export of such seafood as shrimp, squid, crab, and lobster has become a growing source of foreign exchange. There also has been an increase in the number of commercial shrimp farms. The most important freshwater fisheries are located on the plains of the Mekong and Bassac rivers.

Forestry is a major industry. Charcoal production is widespread, and a number of factories produce furniture, pulp, and paper. Plywood, lumber, and rattan products also contribute to the economy. Deforestation and soil degradation, however, threaten the viability of the industry, especially because of increasing domestic demand for forest products.

Industry. Following the reunification of the country in 1975, a concerted effort was made to rapidly transform the private, capitalist industry in the south into a state-run sector. Many industrial operations there were nationalized or forced to become joint state-private enterprises. For industry as a whole, the productivity of both capital and labour declined, and gross output slumped. Heavy industry—plagued by waste and inefficiency, lack of spare parts and raw materials, energy shortages, and poor quality control—led the decline.

Reform measures in the 1980s included introducing incentives, reducing subsidies to inefficient state-run operations, and gradually allowing limited market mechanisms. Light industry registered significant gains, while heavy industry responded more sluggishly but showed some improvement. Private enterprise and collectives grew somewhat at the expense of the state sector.

Mining has remained important, especially in the north, although not to the extent it once was. Coal, especially anthracite, has long been an important export. Phosphate, iron, tin, antimony, and chromium are produced for domestic industry and for export. Gold is used to make jewelry. Manganese, bauxite, lead, tungsten, zinc, and lime also are mined in significant quantities. The fledgling petroleum industry has grown steadily since oil extraction began in 1986.

Food processing is the largest industrial activity in Vietnam. Seafood is processed for export, while coffee and tea are processed both for export and for domestic consumption. Beverages and a variety of condiments also are produced in significant quantities. Textiles are of increasing importance; silk production revived in the 1990s after a period of decline.

Vietnam long has been a major producer of cement. The chemical industry has been growing, with fertilizer being its most important product. Steel is a major part of Vietnam's heavy industry. Rubber is processed for export and is used in domestic production, mainly to make bicycle tires and tubes but also in the manufacture of a wide range of other goods. Other important industrial products include bicycles, machine tools, electrical equipment, celluloid and paper, and leather goods. Electric power generation has increased significantly since 1980, with much of the growth coming from the development of hydroelectricity.

Finance. In the north until 1975 and throughout Vietnam thereafter, the state-owned National Bank of Vietnam (renamed the State Bank of Vietnam) functioned as a government monopoly in the banking sector. Commercial banking facilities were virtually nonexistent. Following economic reforms in the late 1980s, however, this structure was inadequate to attract badly needed foreign trade and investment. An import-export bank was established in 1989 to promote investment and to facilitate trade transactions. As foreign investment gradually increased in the early 1990s, some foreign commercial banks were allowed to establish branch offices in Vietnam, which greatly expanded the scope of banking services available.

Banking reforms

Trade. Both parts of Vietnam experienced trade deficits during the war, and deficits continued after reunification. A trade embargo imposed by the Americans exacerbated problems of low efficiency and poor quality control that hampered exports. In the first decade after reunification, the value of exports was only one-third that of imports. The Soviet Union and the communist countries of eastern Europe came to be Vietnam's most important trading partners.

Vietnam's efforts to increase trade with capitalist countries as part of its larger program of economic reforms took on added urgency with the breakup of the Soviet Union and the demise of the communist governments in eastern Europe in the late 1980s and early 1990s, because trade with these areas was drastically reduced. Subsequently, such countries as Japan, Singapore, and Thailand became major trading partners. Exports increased, and the trade deficit narrowed.

Mineral fuels and lubricants, motor vehicles, machinery, and foodstuffs account for most of Vietnam's imports. Exports consist mainly of primary commodities (*e.g.*, rubber, coal, forest products, and coffee), handicrafts, and such light manufactures as textiles and footwear. In addition, there has been a significant rise in the export of processed seafood and petroleum.

Transportation. The topography of Vietnam renders land transportation between the north and the south difficult, with traffic limited to the narrow coastal corridor. Hanoi and Ho Chi Minh City are connected by rail and highway through this corridor. The nation's road network is extensive but in generally poor condition. The two large

deltas, where most of the population is concentrated, rely heavily on a vast network of navigable rivers and canals.

Ho Chi Minh City and Hanoi have international airports. In addition, a number of smaller cities are connected by domestic air routes. Vietnam's major ports are at Haiphong in the north and Ho Chi Minh City in the south, followed by Da Nang in central Vietnam. There are several other good ports, including Cam Ranh, a superb natural harbour developed extensively by the Americans during the war.

Administration and social conditions. *Government.* The first constitution of the Socialist Republic of Vietnam was adopted in 1980; it was superceded by a second constitution, promulgated in 1992. In addition to reforming Vietnam's government and political structure, the 1992 constitution also outlined major shifts in foreign policy and economic doctrine. In particular, it stressed the development of all economic sectors, including private enterprise, and it granted foreign investors the right to legal ownership of their capital and assets while guaranteeing that their property could not be nationalized by the state.

The 1992 constitution

A unicameral, popularly elected National Assembly is the supreme organ of the government. It elects the president, who is head of state, and the vice president. The cabinet consists of the prime minister, who is nominated by the president and approved by the National Assembly, and deputy prime ministers and the heads of government ministries and various state organizations, who are named by the prime minister and confirmed by the Assembly. The cabinet coordinates and directs the ministries and various state organizations of the central government and supervises the administrative committees at the local government level.

The responsibilities of the ministries usually are divided along narrow functional lines; there are, for example, numerous economic ministries concerned with agriculture and the food industry, marine products, forestry, and water conservancy. Larger ministries tend to be relatively self-sufficient, with their own colleges, training institutions, and health, social, and cultural facilities. There also are several commissions under the cabinet, such as the State Planning Commission. The prime minister's office oversees a number of general departments beneath the ministerial level and committees that are formed to supervise major projects which involve more than one ministry.

The country is divided administratively into 50 provinces and 3 municipalities (Hanoi, Haiphong, and Ho Chi Minh City). These are further subdivided into about 500 districts. At the provincial and district levels, the highest government authority is an elected People's Council, the actual work of which is carried out by administrative committees elected by the councils. Village administration is represented by village People's Councils.

Both the 1980 and 1992 constitutions institutionalized the Vietnamese Communist Party as the sole source of leadership for the state and society. The 1992 document, however, delegated much more authority to the president and to the cabinet (which superceded the earlier Council of Ministers); they were given the task of running the government, while the party became responsible for overall policy decisions. These changes reduced considerably the role of the party. Notably affected were the Politboro and the larger Central Committee—which previously had been the major decision-making bodies of both the party and the state—and the Secretariat and its presiding general secretary—which, in their role of operating the party organization and carrying out the resolutions of the Central Committee and the Politboro, had effectively governed the country.

Popular associations

Nonetheless, the party remains the dominant political institution within Vietnam. Numerous popular associations disseminate party policies and serve as training grounds for potential party members. The Vietnamese Women's Union is an important and active organization. The Ho Chi Minh Communist Youth Union is largely responsible for the Vietnam Youth Federation, while local party units and agricultural cooperative organizations assume leadership over the Farmers' Federation. The Vietnam Federation of Trade Unions has the responsibility of safeguarding workers' welfare, but it does not function as a Western-style bargaining unit.

Justice. The judicial system consists of the courts and the People's Organs of Control. The National Assembly supervises the work of the Supreme People's Court, which is the highest court of appeal and the court of first instance for special cases (such as treason). This court, in turn, supervises the judicial work of local People's Courts, which are responsible to their corresponding People's Councils. The People's Courts function at all levels of government except the village, where the village administrative committee functions as a primary court.

The People's Organs of Control act as watchdogs for the state: they monitor the performance of government agencies, maintain vast powers of surveillance, and act as prosecutors before the People's Courts. The Supreme People's Organ of Control is responsible only to the Standing Committee of the National Assembly.

Armed forces. Military forces include the army, paramilitary regional and provincial forces, the militia, and the reserves. There are separate military commands in Hanoi, Haiphong, and Ho Chi Minh City. Vietnam has maintained a proportionally large military force.

Education. The Vietnamese, with their Confucian traditions, have always placed great importance on education. Rural education in the south was badly disrupted during the war years, and all religious and private schools were nationalized after 1975. The government subsequently pursued a policy of education reform. Nine years of schooling are mandatory and are divided into five years of primary and four years of lower-secondary school. Continuing students are enrolled either in an academic or a vocational upper-secondary program, which lasts three years. Opportunities for advanced education are limited. The University of Hanoi, founded originally by the French and refounded in 1956, is the country's oldest institute of higher education.

Literacy rates are high. Emphasis is placed on training in science and technology, although a lack of equipment hinders the program. Several thousand students are sent abroad each year to study languages and technology. While most students once went to the Soviet Union and the countries of eastern Europe, increasing numbers are now studying in Western countries (including the United States) or in Japan.

Health and welfare. Before reunification, health services were underdeveloped in the rural areas of the south but were well-developed in the north. After 1975 there was a general increase in health facilities and personnel. The health care system is one of the socialist state's greatest achievements; like all other programs in Vietnam, however, it has been severely hampered by a lack of funds since the late 1970s.

Cultural life. Chinese influence permeated all aspects of traditional Vietnamese culture, while Western influences have been strong in the 20th century.

Daily life. Vietnam's Confucian heritage is seen in the importance of family to the Vietnamese. Families are essentially patrilineal, but Vietnamese women work alongside men in many jobs and play a major role in raising children and managing family finances. When possible, the Vietnamese prefer to work from early morning until early evening, with an extended rest period in the heat of midday. In rural areas, both men and women wear trousers and shirts or blouses. On formal occasions and in urban areas, Western-style clothing is common, including skirts and blouses for women. Women still sometimes wear a form of the traditional *ao dai*, a long, slit tunic worn over pants.

Cuisine

Rice is the staple food. Vietnamese cuisine incorporates elements of both Chinese cooking and the cuisines of other Southeast Asian countries. Noodle soup with chicken or beef broth (*pho*), a distinctive kind of spring roll (*cha gio*), and the use of fermented fish sauce (*nuoc mam*) for dipping and seasoning are among the many noteworthy dishes.

The arts. Early Vietnamese poetry was written exlusively in Chinese until the end of the 13th century. By the 15th century, however, a demotic script called *chu nom*

("southern writing") had evolved into a vehicle for writing in vernacular Vietnamese. The Chinese heritage of the elite merged with local oral tradition, producing a truly national literature. A distinctly Vietnamese long narrative poem in verse developed, culminating in the masterpiece of Vietnamese literature, *Kim Van Kieu* (*The Tale of Kieu*), by Nguyen Du (1765–1820). In the 20th century, Vietnamese literature came to be written in a romanized script. In the 1930s a modern Vietnamese literature developed under French influence, featuring poetry, novels, and short stories. Between 1954 and 1975 a cosmopolitan literature stressing creativity and individual freedom flourished in the south, while a state-sponsored literature of Socialist Realism was promoted in the north. After 1975 Socialist Realism became a national orthodoxy, although in the 1980s literature became more lively and diverse in content.

Under communist rule the theatre has been strictly controlled, and all professional performers and other technical staff have become employees of the state. The indigenous *cai luong*, a satirical musical comedy genre that emerged in the south in the early 20th century, is still enormously popular, as are modern plays. There also are theatrical troupes specializing in traditional Chinese opera (called *hat tuong* in the north and *hat boi* in the south), traditional popular operettas (*hat cheo*) of indigenous origin, distinctly Vietnamese water puppetry (*mua roi nuoc*)—in which performances take place on a pool or pond, and water activates the puppets and hides the manipulating apparatus—and circus performances.

Painting has developed slowly and unevenly, bound first by traditional Chinese forms, then by a style imitative of French Impressionism, and more recently by Socialist Realism. High-quality lacquerware, however, continues to be produced. Folk arts persist among the peoples of the central highlands. Women weave blankets and clothing, while men weave baskets and mats. Gongs are the most common of a variety of musical instruments. Crossbows and figures are carved from hardwoods. The Cham and Khmer minorities retain some folk arts, but their traditions seem to be fading.

Sporadic early efforts to develop a film industry in Vietnam have met with little success. Prior to 1975, a small number of undistinguished films were produced in both the north and the south, while the north made a few fine documentaries. Despite continuing financial constraints and technical deficiencies, higher-quality feature films began to appear in the 1980s.

Cultural institutions. Vietnam abounds with a variety of historical sites. Hanoi contains the 11th-century Temple of Literature, the One Pillar Pagoda, and many other ancient sites, as well as the Vietnam History Museum, the National Art Gallery, and the National Library. Its recent past is amply illustrated in the Vietnam Revolution Museum, the People's Army Museum, and a large complex that includes Ho Chi Minh's mausoleum, the house he lived in as president, and the Ho Chi Minh Museum.

Hue

Hue and its environs contain the royal citadel of the last dynasty and numerous royal mausoleums and tombs, as well as many Buddhist pagodas. Ho Chi Minh City has a noteworthy zoo and botanical garden on the edge of the downtown area.

Recreation. Soccer (association football) is exceedingly popular in Vietnam, and volleyball, badminton, wrestling, bicycling, chess, and dominoes are also widely enjoyed. Urban Vietnamese stroll in great numbers on evenings and weekends, especially in the parks and along the banks of lakes and rivers in Hanoi and Ho Chi Minh City. In the larger cities, some young people enjoy Western dancing and listening to Western and modern Vietnamese music in coffeehouses.

Press and broadcasting. Radio, television, and some newspapers and journals are owned and operated by the state. The publishing of newspapers, magazines, and books is regulated by the government, but the strict controls of earlier years were somewhat relaxed during the 1980s. The circulation of many books published in the south between 1954 and 1975, however, is still forbidden.

(G.C.H./N.L.J.)

For statistical data on the land and people of Vietnam, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Origins of the Vietnamese people. Relatively little is known about the origins of the Vietnamese. They first appeared in history as the so-called "Lac" peoples, who lived in the Red River delta region, in what is now northern Vietnam. Some scholars have suggested that the Lac were closely related to other peoples, known as the Viet (called the Yüeh by the Chinese), who inhabited the coastal region of East Asia from the Yangtze River to the Red River delta during the 1st millennium BC. Others have expressed doubt about this supposition, noting that modern-day Vietnamese share many cultural and linguistic traits with other non-Chinese peoples living in neighbouring areas of Southeast Asia. It is now generally believed that the Lac peoples were the result of a mixture between Australo-Melanesian inhabitants who had lived in the area since Paleolithic times and Asiatic peoples who later migrated into the area from China.

Linguistic research, which offers a relatively reliable way of distinguishing the various ethnic groups of Southeast Asia, supports the mixed ethnic and cultural origin of the Vietnamese people. Although the Vietnamese language is distinct, it nevertheless can be described as a fusion of Mon-Khmer, Tai, and Chinese elements. From the monotonic Mon-Khmer language family, Vietnamese derived many of its basic words; from the Tai languages, it took tonality and a number of grammatical elements; and from the Chinese, who at that time were somewhat more culturally advanced than the peoples of the Red River delta, it acquired not only a script but also most of its political, literary, philosophical, and technical vocabulary.

Ethnographic study also reveals the degree to which ancient Vietnamese culture was a composite of elements found among many other peoples within the region. Totemism, animism, tattooing, the chewing of betel nuts, teeth blackening, and many marriage rituals and seasonal festivals indicate the relationship between the Vietnamese and the neighbouring peoples in Southeast Asia. Although Chinese civilization later became the main force in shaping Vietnamese culture, the failure of the Chinese to assimilate the Vietnamese people underscores the fact that strong elements of an authentic local culture must have emerged in the Red River valley long before China established its millennium of rule over Vietnam.

Earliest known history of Vietnam. *Legendary kingdoms.* According to the most authoritative legends, the history of the Vietnamese people begins with King De Minh, a descendant of a divine Chinese ruler who was also the legendary father of Chinese agriculture. De Minh and an immortal fairy of the mountains produced Kinh Duong, ruler of the Land of Red Demons, who married the daughter of the Dragon Lord of the Sea. Their son, Lac Long Quan ("Dragon Lord of Lac"), is regarded as the first authentic Vietnamese king. To make peace with the Chinese, Lac Long Quan married Au Co, a Chinese immortal, who bore him 100 eggs, from which sprang 100 sons. Later, the king and queen separated; Au Co moved with 50 of her sons into the mountains, and Lac Long Quan kept the other 50 sons and continued to rule over the lowlands. Lac Long Quan's eldest son succeeded him as the first of the Hung (or Hong Bang) kings (*vuong*), and he is regarded as the real founder of the Vietnamese nation and of the first Vietnamese dynasty.

Lac Long Quan

This legend and other related legends, most of which received their literary form only after AD 1200, describe in mythical terms the fusion, conflicts, and separation of peoples from the north and south and of peoples from the mountains and the coastal lowlands. The legends show the immortals as mountain dwellers, while the people along the coast are represented by the dragon lords—a division found in many legends throughout Southeast Asia. The retreat of Au Co and 50 of her sons into the mountains may well be a mythical record of a separation among the proto-Vietnamese in the Red River delta: those who left the lowlands could be the ancestors of the Muong, who

still live in the hills surrounding the delta and who are the only ethnic minority of Vietnam closely related in language and customs to the Vietnamese.

According to legend, the Hung dynasty had 18 kings, each of whom ruled for about 150 years. Their country, called Van Lang ("Land of the Tattooed Men"), is said to have included not only the Red River delta but also much of southern China. The last of the Hung kings was overthrown in 258 or 257 BC by a neighbouring warlord, Thuc Phan, who invaded and conquered Van Lang, united it with his kingdom, and called the new state Au Lac, which he then ruled under the name An Duong. Au Lac existed only until 207 BC, when it was incorporated by a former Chinese general, Trieu Da (Chao T'o in Chinese), into the kingdom of Nam Viet (Nan Yüeh in Chinese).

Nam Viet. This kingdom covered much of southern China and was ruled by Trieu Da from his capital near the present site of Canton. Its population consisted chiefly of the Viet who had earlier been driven by the Chinese from their kingdoms south of the Yangtze River. Trieu Da, after throwing off Chinese sovereignty and killing all officials loyal to the Chinese emperor, adopted the customs of the Viet and made himself the ruler of a vast non-Chinese empire. After it had incorporated Au Lac, Nam Viet included not only the Red River delta but also the coastal lands as far south as modern-day Da Nang. The end of Au Lac in 207 BC marks the end of legendary history and the beginning of Vietnamese history, as recorded in Chinese historical annals.

After almost 100 years of diplomatic and military duels between the Han Chinese empire and Trieu Da and his successors, Nam Viet was conquered (111 BC) by the Chinese under the Han emperor Wu-ti. Thus, the territories occupied by the ancestors of the Vietnamese fell under Chinese rule. Nam Viet became the Chinese province Giao Chi (later Giao Chau), which was divided into nine military districts. The three southernmost of these covered the northern half of what is now Vietnam.

Chinese rule

Early society. When China extended its rule over Vietnam, the people of the Red River delta were in transition from the Bronze to the Iron Age, although some stone implements were also still in use. These ancestors of the Vietnamese were already experienced at cultivating rice. They had learned how to irrigate their rice fields by using the tides that backed up the rivers. Plows and water buffalo were still unknown (the land was prepared for cultivation with polished stone hoes), but the proto-Vietnamese are thought to have been able to produce two rice crops annually. They supplemented their diet by fishing and hunting. Their weapons were mainly bows and arrows; the bronze heads of their arrows often were dipped in poison to facilitate killing such larger animals as elephants, whose tusks were traded for iron from China.

The social organization of the early Vietnamese, before Chinese rule, was hierarchical, forming a kind of feudal society that until the mid-20th century existed among the Tai and Muong minority populations of northern Vietnam. Power was held by tribal chiefs at the head of one or several communities. These chiefs were civil, religious, and military leaders, and their power was hereditary; they were large landowners who kept the mass of the people in virtual serfdom. At the head of this aristocracy stood the king, probably the most powerful of the tribal chiefs.

Religion was characterized by the kind of animistic beliefs in supernatural beings and spirits that are common among preliterate agricultural and hunting peoples. Some of the spirits were those of dangerous animals, while others were of deceased important persons who needed to be propitiated. A great religious festival, almost a carnival, was held at the beginning of spring and was marked by abandon and promiscuity.

In all these respects, the inhabitants of the Red River delta, prior to their subjugation by the Chinese, showed numerous affinities with most of the people of mainland and island Southeast Asia. It was not until several centuries after the imposition of Chinese rule that the Vietnamese developed more distinct ethnic characteristics.

Vietnam under Chinese rule. The history of the Vietnamese people under more than a millennium of Chinese rule reveals an evolution toward national identity, which apparently came about as the result of three developments. The first of these was the introduction into the Red River delta of the more advanced civilization of China, including technical and administrative innovations and the more sophisticated level of Chinese learning, which made the Vietnamese the most advanced people of mainland Southeast Asia. The second was the efforts of the Chinese governors to achieve complete Sinicization through the imposition of Chinese culture, customs, and political institutions. The third and most significant development during this period was the resistance of the Vietnamese people to total assimilation and the use they made of the benefits derived from Chinese civilization in their struggle against Chinese political rule.

Soon after extending their domination over what is now northern Vietnam, the Chinese constructed roads, waterways, and harbours to facilitate communications within the region and to ensure that they maintained administrative and military control over it. They improved local agriculture by introducing better methods of irrigation and the use of metal plows and draft animals. They brought with them new tools and weapons, advanced the art of pottery, and used new mining techniques. For more than a century after annexing Nam Viet, however, the Chinese abstained from interfering with the local administration. In the Chinese province of Giao Chau, the hereditary lords exercised control over the peasant population, just as they had done while Giao Chau was a province of Nam Viet. Thus, although Vietnam was divided into military districts headed by Chinese governors, it remained, in fact, a leniently governed Chinese protectorate.

Chinese building programs

This form of government changed in the 1st century AD, when an energetic governor realized that the sway of the local Viet lords over the population was an obstacle to Sinicization. The desire to exploit the fertile Red River delta and its mountainous backcountry was certainly one reason why the expansionist Han dynasty wanted to hold on to Vietnam: there were vast forests and precious metals in the mountains, pearls in the sea, elephants with tusks of ivory, and a peasantry that could be taxed and recruited for forced labour. China's main interest in holding the Red River delta, however, was its value as an important stopover for ships engaged in the Han dynasty's nascent maritime trade with the East Indies, India, and even the Middle East. Vessels from many countries with which China developed commercial relations docked at the harbours of the Vietnamese coast, not only bringing new goods but also establishing contacts with a wider world and thus promoting the development of the country. In this process, which began early in the 1st century AD, economic, cultural, and political functions developed that the hereditary local lords were unfit to discharge—another reason why direct Chinese rule through the importation of an increasing number of Chinese officials became necessary.

As in all regions conquered by the Chinese Han dynasty (206 BC–AD 221, with a brief interruption in AD 8–23), efforts to set up direct Chinese rule were accompanied by a variety of attempts to transform the people of the Red River delta into Chinese. Local customs were suppressed, and Chinese customs, rites, and institutions were imposed by force. Taoist and Confucian teachings were pressed, together with instruction in the Chinese language; even Chinese clothing and hairstyles became obligatory. Many of these Chinese innovations were beneficial to the Vietnamese and were readily integrated into the indigenous local culture, but Sinicization never succeeded in reconciling the Vietnamese people, especially their leaders, with Chinese political domination. Not only the masses of the people but even the educated Vietnamese who knew Chinese and wrote only in Chinese held on to the local spoken language.

Attempts at Sinicization

The first major rebellion against Chinese rule broke out in AD 40, led by the noblewoman Trung Trac, whose husband, a tribal lord, had been executed by the Chinese. She and her sister, Trung Nhi, gathered together the tribal chiefs and their armed followers, attacked and overwhelmed the Chinese strongholds, and had them-

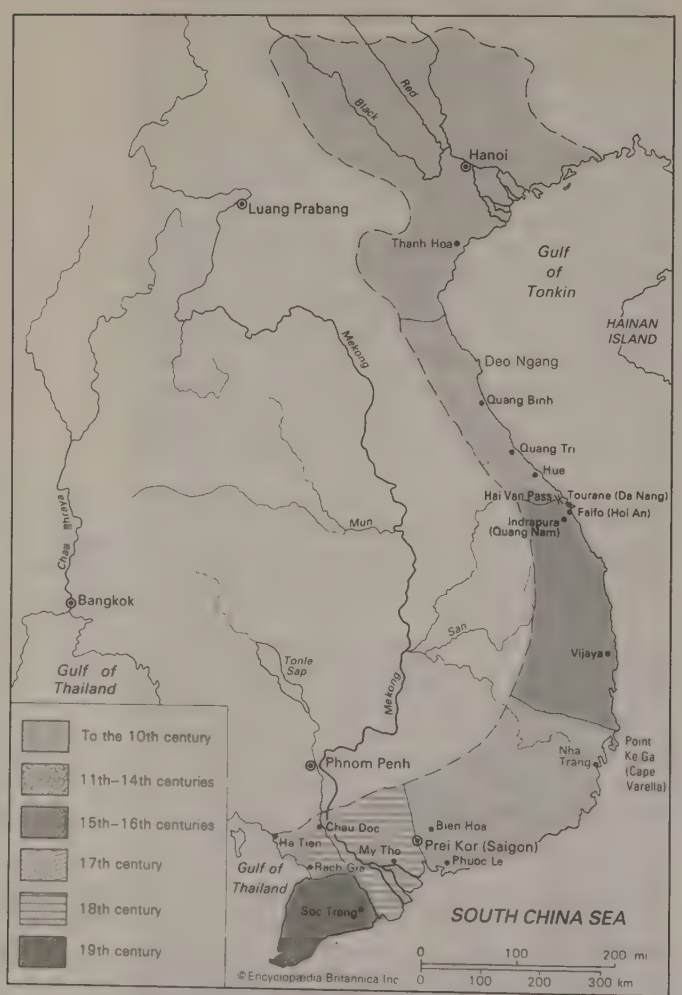
Vietnamese resistance

selves proclaimed queens of an independent Vietnamese kingdom. Three years later a powerful army sent by the Han emperor reestablished Chinese rule; the local aristocracy was deprived of all power, Vietnam was given a centralized Chinese administration, and Sinicization was resumed with increased intensity. The Trung sisters apparently were put to death by their conquerors.

Chinese rule, although challenged several more times, remained secure so long as China itself was effectively controlled by its own emperors. When the T'ang dynasty (618–907) fell into decay in the early 10th century, a series of uprisings broke out in Vietnam, which led in 939 to the restoration of Vietnamese independence.

The first period of independence. *The Ly dynasty.* Ngo Quyen, the Vietnamese commander who had defeated the Chinese in 939, became the first head of the new state of Vietnam. For more than a half century, however, independence brought neither peace nor political stability to Vietnam. In the early 11th century, Vietnam finally was brought together under a centralized administration by Ly Thai To, the founder of the Ly dynasty (sometimes called the Later Ly dynasty; 1009–1225). The Ly rulers established their capital at Hanoi, in the heart of the Red River delta, modernized the agricultural system, and replaced the divisive local lords with a system of state officials trained in a civil service institute set up on the Chinese model in 1076.

Although the new state, called Dai Viet, made considerable political, economic, and cultural progress, it soon encountered problems with its neighbour to the south, the Islāmic, Indianized state of Champa on the central coast. Dai Viet and Champa fought several wars in the 12th and 13th centuries. Dai Viet also clashed with the Khmer (Cambodian) state of Angkor, then the greatest power in mainland Southeast Asia.



Expansion of precolonial Vietnam.

The Tran dynasty. By then, the Ly dynasty was already in a state of decline. It was succeeded, after a period of civil strife, by a new dynasty called the Tran, which reigned from 1225 to 1400. For most of their rule, the Tran kings pursued the same policies that had made the country strong under the Ly; the Tran continued to clash with Champa, but they also were able to maintain several periods of peaceful coexistence. The primary challenge to Vietnamese independence, however, came from the north. The Yüan (Mongol) dynasty, which had come to power in China in 1279, sent armies estimated at more than 300,000 soldiers to restore the Red River delta to Chinese rule. The Tran resisted stubbornly and eventually were able to drive out the invaders. The general who commanded the Vietnamese forces, Tran Hung Dao, is still venerated as one of the great heroes of Vietnamese history.

The drain of these wars on Vietnam's resources, together with the declining vigour of its rulers, brought on a deep economic and social crisis and the overthrow of the Tran dynasty in 1400. The deposed Tran ruler appealed to China to help him regain the throne. China, by then ruled by the Ming dynasty (1368–1644), readily complied with the request, and China again invaded Vietnam in 1407. The Ming set up a direct Chinese administration, and these officials resumed the policies of assimilation begun by their imperial predecessors.

Chinese invasion of 1407

Expansion, division, and reunification. By the beginning of the 15th century, the cultural evolution of the Vietnamese people had reached a point at which any attempt to make them Chinese could only strengthen their nationalist sentiments and arouse their determination to throw off the Chinese yoke. Under the leadership of Le Loi, a wealthy landowner in the province of Thanh Hoa south of the Red River delta, a movement of national resistance started in 1418; after a 10-year struggle, the Chinese were forced to evacuate. Le Loi, who ascended the throne shortly thereafter under the name of Le Thai To, became the founder of the third great Vietnamese dynasty, the Later Le (sometimes simply referred to as the Le). Although the Later Le was not actually in power after 1600, it nominally headed the state until 1788.

The Later Le dynasty. Like the better rulers of the Ly and Tran dynasties, Le Thai To and some of his successors introduced many reforms. They gave Vietnam the most advanced legal code in Southeast Asia; promoted art, literature, and education; advanced agriculture; protected communal lands against the greed of large landowners; and even enforced a general redistribution of land among the entire population at the expense of the large landowners. The problem of the landless remained acute, however, because of population increases and the limited amount of available land in the north. The lack of land was one of the reasons the Le dynasty pursued a policy of territorial expansion, and it was a chief motive behind their efforts to drive the Chams from the small but fertile deltas to the south. Most of Champa was conquered in 1471 under the leadership of Le Thanh Tong (ruled 1460–97). Soldiers in the advancing Vietnamese army were settled in newly established villages south from the vicinity of Da Nang to the neighbourhood of Nha Trang, in what became the first great Vietnamese push into the south. The elimination of Champa was followed by incursions into the Cambodian territory of the Mekong delta, which the declining Khmer empire no longer was able to defend. Saigon became Vietnamese shortly before 1700, and the rest of the south followed during the next 60 years. With the exception of the southern province of Soc Trang, which was not annexed until 1840, Vietnam had reached its present size by 1757.

This extension of Vietnam to a length of some 1,000 miles altered the historical evolution of the state: up to that point, Vietnam's chief characteristic had been the existence of a strong central power at the head of a unified administration. The country subsequently was divided twice, and its partitioned governments were at war with each other for decades.

Two divisions of Vietnam. The first and shorter division of the country occurred soon after the elimination of Champa. The governor of Hanoi, Mac Dang Dung, made himself master of Vietnam in 1527. The deposed Le rulers

and the generals loyal to them regained control of the lands south of the Red River delta in 1545, but only after nearly 50 years of civil war could they reconquer Hanoi and the north.

Of much longer duration and greater historical significance was the second division of Vietnam, which occurred about 1620, when the noble Nguyen family, who had governed the country's growing southern provinces from Hue since 1558, rejected Hanoi's suzerainty. In Hanoi the Le monarchs were rulers in name only after the country was reunited following its first division; all real power was in the hands of the Trinh family, who had made themselves hereditary princes in charge of the government. For 50 years the Trinh rulers tried in vain to regain control of the southern half of the country by military means. The failure of their last campaign in 1673 was followed by a 100-year truce, during which both the Nguyen and the Trinh paid lip service to Vietnamese unity under the Le dynasty but maintained separate governments over the two halves of the country.

The Tay Son rebellion

National unity was reestablished only after a 30-year period of revolution, political chaos, and civil war (1772–1802). Although the revolution started in the south, it was directed against the ruling houses of both south and north. It was led by three brothers, whose name in history—Tay Son—was that of their native village. The Tay Sons overthrew the southern regime in 1777 and killed the ruling family. While the Tay Sons waged war against the north, one member of the southern royal family—Nguyen Anh, who had escaped the massacre—regained control of Saigon and the deep south in 1778, but he was driven out again by the Tay Sons in 1783. When the Tay Sons also defeated the Trinh in 1786 and occupied Hanoi, Vietnam was briefly reunited under Tay Son rule. In 1788 the Chinese tried to exploit the crisis in Vietnam, but the Tay Son rulers—who had abolished the Later Le dynasty—were able to defeat the Chinese invaders. During that same year, however, Nguyen Anh succeeded, with French military assistance, in occupying Saigon and the Mekong delta. In a series of campaigns that lasted 14 years, Nguyen Anh defeated the Tay Sons and gained control of the entire country. When Hue and Hanoi fell to his armies in 1802, he proclaimed himself emperor, under the name Gia Long, of a reunited Vietnam.

State and society of precolonial Vietnam. The rule of Gia Long and his successors until the conquest of Vietnam by France in the late 19th century brought no innovations in the organization of the state, the basic character of which already had been firmly established by the Ly emperors during the 11th century. The Ly had successfully fought the revival of a local feudalism, which was rooted in the powers exercised by the tribal chiefs before the coming of the Chinese. From the 11th century on, Vietnam remained a centralized state headed by a monarch whose absolute powers were said to derive from a mandate from heaven—one aspect of the thoroughly Confucian character of the Vietnamese state. The Ly established a fixed hierarchy of state officials that followed the Chinese model; it consisted of nine degrees of civil and military mandarins who were appointed by the emperor and were responsible to only him. All mandarins—those at the very top at the imperial court as well as those in the lowest ranks of the provincial and local administration—were recruited in only one way: through civil service examinations taken after years of study. As a rule, only the wealthy could spend the time required for these studies. Nevertheless, except in periods of dynastic decline when offices sometimes were for sale, the road to positions of power was through scholarship, not wealth.

The concept of a division of powers was alien to the precolonial rulers. The emperor, with the help of high court mandarins, was not only the supreme lawmaker and head of all civil and military institutions but also the dispenser of justice in both criminal and civil cases, and he delegated his powers to the hierarchy of mandarins in the provinces and villages. Even public functions of a religious character were the sole prerogative of the emperor and his representatives in the lower levels of the administration. No military caste ever exercised control over the state, no religious hierarchy existed outside the mandarins, and no aristocracy with political influence was allowed to arise. Titles of nobility, bestowed as honours, were not hereditary.

Economic policies

The economic policies of the great Vietnamese dynasties also favoured the maintenance of imperial and mandarin power. Through the 900 years of independence between Chinese domination and French colonial rule, the country's economy remained almost exclusively agricultural. Artisan and fishing villages existed, and there was some mining; but the mass of people were engaged in the cultivation of rice, and neither national nor international trade was systematically promoted. No property-owning middle class of merchants ever threatened the authority of the scholar mandarins, and the periodically rising power of great landowners was diminished from time to time through the redistribution of land. Gia Long and his successor, Minh Mang, actually abolished all huge landholdings during the first half of the 19th century. Theoretically, the emperor owned all the land, and it was by imperial decree that the settlers on newly conquered territories received their plots in the villages that sprang up south from the Red River delta to the Mekong delta.

Vietnam's rigid absolutism was limited to a certain extent by the Confucian concept that held the family to be the basic unit of civilized society; submission to the authority of the family head thus was the foremost moral obligation of every citizen. The autocratic character of society also was eased slightly by the limited authority granted to the village administration, whose purely local affairs were handled by a council of notables elected, as a rule, from the more prosperous or otherwise prominent citizens. Among the duties of these notables were the enforcement of law, the conscription of army and forced-labour recruits, and the assessment of taxes. Next to devotion to family, loyalty to the village was traditionally the first duty of every Vietnamese.

Western penetration into Vietnam. In 1516 Portuguese adventurers arriving by sea inaugurated the era of Western penetration into Vietnam. They were followed in 1527 by visiting Dominican missionaries, and eight years later a Portuguese port and trading centre was established at Faifo (modern Hoi An), south of present-day Da Nang. More Portuguese missionaries arrived later in the 16th century, and they were followed by other Europeans. The best-known of these was the French Jesuit missionary Alexandre de Rhodes, who completed a transcription of the Vietnamese language into roman script that later was adopted by modern Vietnamese as their official writing system, *quoc ngu* ("national language").

By the end of the 17th century, however, the two rival Vietnamese states had lost interest in maintaining relations with European countries; the only window left open to the West was at Faifo, where the Portuguese retained a trading mission. For decades the French had tried without success to retain some influence in the country. Only at the end of the 18th century was a missionary named Pigneau de Béhaine able to restore a French presence by assisting Nguyen Anh in wresting control of the country from the Tay Sons.

Vietnamese actions against the French

Upon becoming emperor, however, Nguyen Anh (now Gia Long) did not favour Christianity. Under his strongly anti-Western successor, Minh Mang (ruled 1820–41), all French advisers were dismissed, while seven French missionaries and an unknown number of Vietnamese Christians were executed. After 1840 French Roman Catholic interests openly demanded military intervention to prevent the persecution of missionaries. In 1847 the French took reprisals against Vietnam for expelling additional missionaries, but 10 years passed before Paris prepared a military expedition against Vietnam.

The conquest of Vietnam by France. The decision to invade Vietnam was made by Napoleon III in July 1857. It was the result not only of missionary propaganda but also, after 1850, of the upsurge of French capitalism, which generated the colonial concept of a need for overseas markets and the collateral request for a larger French share in Asian territories conquered by the West. The naval commander in East Asia, Rigault de Genouilly, long

an advocate of French military action against Vietnam, was ordered to attack the harbour and city of Tourane (modern Da Nang) and to turn it into a French military base. Genouilly arrived at Tourane in August 1858 with 14 vessels and 2,500 men; the French stormed the harbour defenses on September 1 and occupied the town a day later. Genouilly soon recognized, however, that he could make no further progress around Tourane and decided to attack Saigon. Leaving a small garrison behind to hold Tourane, he sailed southward in February 1859 and seized Saigon two weeks later.

Vietnamese resistance prevented the French from advancing beyond Saigon, and it took French troops, under new command, until 1861 to occupy the three adjacent provinces. The Vietnamese, unable to mount effective resistance to the invaders and their modern weapons, concluded a peace treaty with France in June 1862, which ceded the conquered territories to the latter. Five years later additional territories in the south were placed under French rule. The entire colony was named Cochinchina.

It had taken the French slightly more than eight years to make themselves masters of Cochinchina (a protectorate already had been imposed on Cambodia in 1863). It took them 16 more years to extend their control over the rest of the country. They made a first attempt to enter the Red River delta in 1873, after a French naval officer and explorer named Francis Garnier had shown, in a hazardous expedition, that the Mekong River could not serve as a trade route into southwestern China. Garnier had some support from the French governor of Cochinchina, but when he was killed in a battle with Chinese pirates near Hanoi, the attempt to conquer the north collapsed.

Within a decade, France had returned to the challenge. In April 1882, with the blessing of Paris, the administration at Saigon sent a force of 250 men to Hanoi under Captain Henri Rivière. When Rivière was killed in a skirmish, Paris moved to impose its rule by force over the entire Red River delta. In August 1883 the Vietnamese court signed a treaty that turned northern Vietnam (named Tonkin by the French) and central Vietnam (named Annam, based on an early Chinese name for the region) into French protectorates. Ten years later the French annexed Laos and added it to the so-called Indochinese Union, which the French created in 1887. The union consisted of the colony of Cochinchina and the four protectorates of Annam, Tonkin, Cambodia, and Laos.

The Indochinese Union

Colonial Vietnam. *French administration.* The French now moved to impose a Western-style administration on their colonial territories and to open them to economic exploitation. Under Governor-General Paul Doumer, who arrived in 1897, French rule was imposed directly on all levels of administration, leaving the Vietnamese bureaucracy without a trace of real power. Even Vietnamese emperors were deposed at will and replaced by others willing to serve the French. All important positions within the bureaucracy were staffed with officials imported from France; even in the 1930s, after several periods of reforms and concessions to local nationalist sentiment, Vietnamese officials were employed only in minor positions and at very low salaries, and the country was still administered along the lines laid down by Doumer.

Doumer's economic and social policies also determined, for the entire period of French rule, the development of French Indochina, as the colony became known in the 20th century. The railroads, highways, harbours, bridges, canals, and other public works built by the French were almost all started under Doumer, whose aim was a rapid and systematic exploitation for the benefit of France of Indochina's potential wealth; Vietnam was to become a source of valuable raw materials and a market for tariff-protected goods of French industries. The exploitation of natural resources for direct export was the chief purpose of all French investments, with rice, coal, rare minerals, and later also rubber as the main products. Doumer and his successors up to the eve of World War II were not interested in promoting industry, the development of which was limited to the production of goods for immediate local consumption. Among these industries—located chiefly at Saigon, Hanoi, and Haiphong—were breweries, distilleries, small sugar refineries, rice and paper mills, and glass and cement factories. The greatest industrial establishment was a textile factory at Nam Dinh, which employed more than 5,000 workers. The total number of workers employed by all industries and mines in Vietnam was 100,000 in 1930. Because the aim of all investments was not the systematic economic development of the colony but the attainment of immediate high returns for investors, only a small fraction of the profits was reinvested.

Effects of French colonial rule. Whatever economic progress Vietnam made under the French after 1900 benefited only the French and the small class of rich Vietnamese created by the colonial regime. The masses of the Vietnamese people were deprived of such benefits by the social policies inaugurated by Doumer and maintained even by his more liberal successors, such as Paul Beau (1902–07), Albert Sarraut (1911–14 and 1917–19), and Alexandre Varenne (1925–28). Through the construction of irrigation works, chiefly in the Mekong delta, the area of land devoted to rice cultivation quadrupled between 1880 and 1930. During the same period, however, the individual peasant's rice consumption decreased without the substitution of other foods. The new lands were not distributed among the landless and the peasants but were sold to the highest bidder or given away at nominal prices to Vietnamese collaborators and French speculators. These policies created a new class of Vietnamese landlords and a class of landless tenants who worked the fields of the landlords for rents of up to 60 percent of the crop, which was sold by the landlords at the Saigon export market. The mounting export figures for rice resulted not only from the increase in cultivable land but also from the mounting degree of peasant exploitation.

Economic inequalities

The peasants who owned their land were rarely better off than the landless tenants. The peasants' share of the price of rice sold at the Saigon export market was less than 25 percent. Peasants continually lost their land to the large owners because they were unable to repay loans given them by the landlords and other moneylenders at exorbitant interest rates. As a result, the large landowners of Cochinchina (less than 3 percent of the total number of landowners) owned 45 percent of the land, while the small peasants (who accounted for about 70 percent of the owners) owned only about 15 percent of the land. The number of landless families in Vietnam before World War II was estimated at half of the population.

The peasants' share of the crop after the landlords, the moneylenders, and the middlemen (mostly Chinese) between producer and exporter had taken their share was still more drastically reduced by the direct and indirect taxes the French had imposed to finance their ambitious program of public works. Other ways of making the Vietnamese pay for the projects undertaken for the benefit of the French were the recruitment of forced labour for public works and the absence of any protection against exploitation in the mines and rubber plantations, although the scandalous working conditions, the low salaries, and the lack of medical care were frequently attacked in the French Chamber of Deputies in Paris. The mild social legislation decreed in the late 1920s was never adequately enforced.

Apologists for the colonial regime claimed that French rule led to vast improvements in medical care, education, transport, and communications. The statistics kept by the French, however, appear to cast doubt on such assertions. In 1939, for example, no more than 15 percent of all school-age children received any kind of schooling, and about 80 percent of the population was illiterate, in contrast to precolonial times when the majority of the people possessed some degree of literacy. With its more than 20,000,000 inhabitants in 1939, Vietnam had but one university, with fewer than 700 students. Only a small number of Vietnamese children were admitted to the lycées (secondary schools) for the children of the French. Medical care was well organized for the French in the cities, but in 1939 there were only 2 physicians for every 100,000 Vietnamese, compared with 76 per 100,000 in Japan and 25 in the Philippines.

Two other aspects of French colonial policy are signifi-

cant when considering the attitude of the Vietnamese people, especially their educated minority, toward the colonial regime: one was the absence of any kind of civil liberties for the native population, and the other was the exclusion of the Vietnamese from the modern sector of the economy, especially industry and trade. Not only were rubber plantations, mines, and industrial enterprises in foreign hands—French, where the business was substantial, and Chinese on the lower levels—but all other business was as well, from local trade to the great export-import houses. The social consequence of this policy was that, apart from the landlords, no property-owning indigenous middle class developed in colonial Vietnam. Thus, capitalism appeared to the Vietnamese to be a product of foreign rule, a fact that, together with the lack of any Vietnamese participation in government, profoundly influenced the nature and orientation of the national resistance movements.

Resistance to French rule

Movements of national liberation. The anticolonial movement in Vietnam can be said to have started with the establishment of French rule. Many local officials of Cochinchina refused to collaborate with the French. Some led guerrilla groups, composed of the remnants of the defeated armies, in attacks on French outposts. A much broader resistance movement developed in Annam in 1885, led by the great scholar Phan Dinh Phung, whose rebellion collapsed only after his death in 1895.

The main characteristic of the national movement during this first phase of resistance, however, was its political orientation toward the past. Filled with ideas of precolonial Vietnam, its leaders wanted to be rid of the French in order to reestablish the old imperial order. Because this aspiration could have little meaning for the generation that came to maturity after 1900, this first stage of anticolonial resistance did not survive the death of its leader.

Modern nationalism. A new national movement arose in the early 20th century. Its most prominent spokesman was Phan Boi Chau, with whose rise the old traditionalist opposition gave way to a modern nationalist leadership that rejected French rule but not Western ideas, science, and technology. In 1905 Chau went to Japan. His plan, mildly encouraged by some Japanese statesmen, was to free Vietnam with Japanese help. Chau smuggled hundreds of young Vietnamese into Japan, where they studied the sciences and underwent training for clandestine organization, political propaganda, and terrorist action. Inspired by Chau's effective writings, nationalist intellectuals in Hanoi opened the Free School of Tonkin in 1907, which soon became a centre of anti-French agitation and consequently was suppressed after a few months. Also, under the inspiration and guidance of Chau's followers, mass demonstrations demanding a reduction of high taxes took place in many cities in 1908. Hundreds of demonstrators and suspected organizers were arrested—some were condemned to death, while others were sent to Poulo Condore (Con Son) Island in the South China Sea, which the French turned into a penal camp for Vietnamese nationalists.

Phan Boi Chau went to China in 1910, where a revolution had broken out against the Ch'ing (Manchu) dynasty. There he set up a republican government-in-exile to attract the support of nationalist groups. After the French arranged his arrest and imprisonment in China (1914–17), however, his movement began to decline. In 1925 Chau was seized by French agents in Shanghai and brought back to Vietnam for trial; he died under house arrest in 1940.

After World War I the movement for national liberation intensified. A number of prominent intellectuals pursued the hope of obtaining political concessions from the colonial regime through collaboration with the French. The failure of such reformist efforts led to a revival of clandestine and revolutionary groups, especially in Annam and Tonkin; among these was the Vietnamese Nationalist Party (Viet Nam Quoc Dan Dang, founded in 1927 and usually referred to as the VNQDD). The VNQDD preached terrorist action and penetrated the garrisons of indigenous troops with a plan to oust the French in a military uprising. On the night of Feb. 9–10, 1930, the troops of only one garrison in Tonkin killed their French officers, but they were overwhelmed a day later and summarily executed. A wave of repression followed that took hundreds of lives and sent thousands to prison camps. The VNQDD was virtually destroyed, and for the next 15 years it existed mainly as a group of exiles in China supported by the Chinese Nationalist Party (Kuomintang).

French repression

Vietnamese communism. For yet another reason, the year 1930 was an important one in the history of Vietnam. Five years earlier, a new figure, destined to become the most prominent leader in the national movement, had appeared on the scene as an exiled revolutionary in South China. He was Nguyen Ai Quoc, better known by his later pseudonym of Ho Chi Minh. In June 1925 Ho Chi Minh founded the Revolutionary Youth League of Vietnam, the predecessor of the Indochinese Communist Party.

Ho Chi Minh had left Vietnam as a young seaman in 1911 and traveled widely before settling in Paris in 1917. He joined the Communist Party of France in 1920 and later spent several years in Moscow and China in the service of the international communist movement. After making his Revolutionary Youth League the most influential of all clandestine resistance groups, he succeeded in early 1930 in forming the Vietnamese Communist Party—from late 1930 called the Indochinese Communist Party—from a number of competing communist organizations. In May 1930 the communists exploited conditions of near starvation over large areas of central Vietnam by staging a broad peasant uprising, during which numerous Vietnamese officials and many landlords were killed, and "Soviet" administrations were set up in several provinces of Annam. It took the French until the spring of 1931 to suppress this movement and, in an unparalleled wave of terror, to reestablish their own control.

Unlike the dispersed and disoriented leadership of the VNQDD and some smaller nationalist groups, the Indochinese Communist Party recovered quickly from the setback of 1931, relying on cadres trained in the Soviet Union and China. After 1936, when the French extended some political freedoms to the colonies, the party skillfully exploited all opportunities for the creation of legal front organizations, through which its influence on intellectuals, workers, and peasants was increased. When political freedoms were again curtailed at the outbreak of World War II, the Communist Party, now a well-disciplined organization, was forced back into hiding.

World War II and independence. For five years during World War II, Indochina was a French-administered possession of Japan. On Sept. 22, 1940, Jean Decoux, the French governor-general appointed by the Vichy government after the fall of France, concluded an agreement with the Japanese that permitted the stationing of 30,000 Japanese troops in Indochina and the use of all major Vietnamese airports by the Japanese military. The agreement made Indochina the most important staging area for all Japanese military operations in Southeast Asia. The French administration cooperated with the Japanese occupation forces and was ousted only toward the end of the war (in March 1945), when the Japanese began to fear that the French forces might turn against them as defeat approached. After the French had been disarmed, Bao Dai, the last French-appointed emperor of Vietnam, was allowed to proclaim the independence of his country and to appoint a Vietnamese national government at Hue, but all real power remained in the hands of the Japanese military commanders.

Japanese ouster of the French

Meanwhile, in May 1941, at Ho Chi Minh's urging, the Communist Party formed a broad nationalist alliance under its leadership called the League for the Independence of Vietnam, which subsequently became known as the Viet Minh. After a short period in jail, Ho was released by the Chinese and began to cooperate with Allied forces by providing information on Japanese troop movements in Indochina. At the same time, he sought recognition of the Viet Minh as the legitimate representative of Vietnamese nationalist aspirations. When the Japanese surrendered in August 1945, the communist-led Viet Minh ordered a general uprising, and, with no one organized to oppose them, they were able to seize power in Hanoi. Bao Dai, the Vietnamese emperor, abdicated a few days later and declared his fealty to the newly proclaimed Democratic Republic of Vietnam.

Clearly the Communist Party had gained the upper hand in its struggle to outmaneuver its disorganized rivals, such as the noncommunist VNQDD. The French, however, were determined to restore their own colonial presence in Indochina and, with the aid of British occupation forces, seized control of Cochinchina. Thus, at the beginning of 1946, there were two Vietnams: a communist north and a noncommunist south. (J.Bu./W.J.D.)

The First Indochina War. Negotiations between the French and Ho Chi Minh led to an agreement in March 1946 that appeared to promise a peaceful solution. Under the agreement France would recognize the Viet Minh government and give Vietnam the status of a free state within the French Union. French troops were to remain in Vietnam, but they would be withdrawn progressively over five years. For a period in early 1946 the French cooperated with Ho Chi Minh as he consolidated the Viet Minh's dominance over other nationalist groups, in particular those politicians who were backed by the Chinese Nationalist Party.

Despite tactical cooperation between the French and the Viet Minh, their policies were irreconcilable: the French aimed to reestablish colonial rule, while Hanoi wanted total independence. French intentions were revealed in the decision of Georges-Thierry d'Argenlieu, the high commissioner for Indochina, to proclaim Cochinchina an autonomous republic in June 1946. Further negotiations did not resolve the basic differences between the French and the Viet Minh. In late November 1946 French naval vessels bombarded Haiphong, causing several thousand civilian casualties; the subsequent Viet Minh attempt to overwhelm French troops in Hanoi in December generally is considered to be the beginning of the First Indochina War.

Outbreak of war

Initially confident of victory, the French long ignored the real political cause of the war—the desire of the Vietnamese people, including their anticommunist leaders, to achieve unity and independence for their country. French efforts to deal with this problem were devious and ineffective. The French reunited Cochinchina with the rest of Vietnam in 1949, proclaiming the Associated State of Vietnam, and appointed the former emperor Bao Dai as chief of state. Most nationalists, however, denounced these maneuvers, and leadership in the struggle for independence from the French remained with the Viet Minh.

Meanwhile, the Viet Minh waged an increasingly successful guerrilla war, aided after 1949 by the new communist government of China. The United States, fearful of the spread of communism in Asia, sent large amounts of aid to the French. But the French were shaken by the fall of their garrison at Dien Bien Phu in May 1954 and agreed to negotiate an end to the war at an international conference in Geneva.

The two Vietnams (1954–65). The agreements concluded in Geneva in April–July 1954 (collectively called the Geneva Accords), which were signed by French and Viet Minh representatives, provided for a cease-fire and for a temporary division of the country into two military zones at latitude 17° N. All Viet Minh forces were to withdraw north of that line, and all French and Associated State of Vietnam troops were to remain south of it; permission was granted for refugees to move from one zone to the other within a given time limit. An international commission was established, composed of Canadian, Polish, and Indian members under an Indian chairman, to supervise the execution of the agreement.

This agreement left the Democratic Republic of Vietnam in control of only the northern half of the country. The last of the Geneva Accords—called the Final Declaration—provided for elections, supervised by the commission, to be held throughout Vietnam in July 1956 in order to unify the country. Viet Minh leaders appeared certain to win these elections, and the United States and South Vietnam would not approve or sign the Final Declaration; elections were never held.

The two Vietnams now began to reconstruct their war-ravaged country. With assistance from the Soviet Union and China, the Hanoi government in the north embarked on an ambitious program of socialist industrialization; they also began to collectivize agriculture in earnest in 1958. In the south a new government appointed by Bao Dai began to build a new country. Ngo Dinh Diem, a Roman Catholic, was named prime minister and succeeded with American support in stabilizing the anticommunist regime in Saigon. He eliminated pro-French elements in the military and abolished the local autonomy of several religious-political groups. Then, in a government-controlled referendum in October 1955, Diem removed Bao Dai as chief of state and made himself president of the Republic of Vietnam.

Diem's early success in consolidating power did not result in concrete political and economic achievements. Plans for land reform were sabotaged by entrenched interests. With the financial backing of the United States, the regime's chief energies were directed toward building up the military and a variety of intelligence and security forces to counter the still-influential Viet Minh. Totalitarian methods were directed against all who were regarded as opponents, and the favouritism shown to Roman Catholics alienated the majority Buddhist population. Loyalty to the president and his family was made a paramount duty, and Diem's brother, Ngo Dinh Nhu, founded an elitist party to clandestinely spy on officials, army officers, and prominent local citizens. Diem also refused to participate in the all-Vietnamese elections described in the Final Declaration. With support from the north, communist-led forces—popularly called the Viet Cong—launched an insurgency movement to seize power and reunify the country. The insurrection appeared close to succeeding, when Diem's army overthrew him in November 1963. Diem and his brother Nhu were killed in the coup.

Rule of Diem

The Second Indochina War. The government that seized power after Diem's ouster, however, was no more effective than its predecessor. A period of political instability followed, until the military firmly seized control in June 1965 under Nguyen Cao Ky. The militant Buddhists who had helped overthrow Diem strongly opposed Ky's government, but he was able to break their resistance. Civil liberties were restricted, political opponents—denounced as neutralists or pro-communists—were imprisoned, and political parties were allowed to operate only if they did not openly criticize government policy. The character of the regime remained largely unchanged after the presidential elections in September 1967, which led to the election of General Nguyen Van Thieu as president.

No less evident than the oppressive nature of the Saigon regime was its inability to cope with the Viet Cong. Aided by a steady infiltration of weapons and advisers from the north, the fighting strength of the insurgent movement grew from about 30,000 men in 1963 to about 150,000 in 1965 when, in the opinion of many American intelligence analysts, the survival of the Saigon regime was seriously threatened. In addition, the political opposition in the south to Saigon became much more organized. The National Front for the Liberation of the South, popularly called the National Liberation Front (NLF), had been organized in late 1960; within four years it had a huge following.

Growing American involvement in the war. Until 1960 the United States had supported the Saigon regime and its army only with military equipment, financial aid, and, as permitted by the Geneva Accords, 700 advisers for training the army. The number of advisers had increased to 17,000 by the end of 1963, and they were joined by an increasing number of American helicopter pilots. All this assistance, however, proved insufficient to halt the advance of the Viet Cong, and in February 1965 U.S. President Lyndon B. Johnson ordered the bombing of North Vietnam, hoping to prevent further infiltration of arms and troops into the south. Four weeks after the bombing began, the United States started sending troops into the south. By July the number of U.S. troops had reached 75,000; it continued to climb until it stood at more than 500,000 early in 1968. Fighting beside the Americans were some 600,000 regular South Vietnamese troops and regional and self-defense forces, as well as smaller contingents from South Korea, Thailand, Australia, and New Zealand.

Buildup of U.S. troops

Three years of intensive bombing of the north and fight-

ing in the south, however, did not weaken the will and strength of the Viet Cong and their allies from the north. The infiltration of personnel and supplies down the famous Ho Chi Minh Trail increased, and some 100,000 regular troops from the north became more important. The continuing strength of the insurgent forces became evident in the so-called Tet Offensive that began in late January 1968, during which the Viet Cong and North Vietnamese attacked more than 100 cities and military bases. After that, the conviction grew in the U.S. government that it was not politically acceptable to sustain the war at such a level, and President Johnson ordered the bombing of the north to be reduced. This decision opened the way for U.S. negotiations with Hanoi, which began in Paris in May 1968. Once all bombing of the North was halted in November, the Paris talks were enlarged to include representatives of the NLF and the Saigon regime.

The war continued under President Richard M. Nixon, who began to withdraw U.S. troops gradually. However, public opposition to the war escalated in 1970 after he ordered attacks on the Ho Chi Minh Trail in Laos and Viet Cong sanctuaries inside Cambodia. In the meantime, the peace talks went on in Paris. (J.Bu./M.E.O./W.J.D.)

Withdrawal of U.S. troops. In January 1973 a peace treaty was signed by the United States and all three Vietnamese parties. It provided for the complete withdrawal of U.S. troops within 60 days and created a political process for the peaceful resolution of the conflict in the south. Nothing was said, however, about the presence of North Vietnamese troops in South Vietnam. The Paris Agreement did not bring an end to the fighting in Vietnam. The Saigon regime made a determined effort to eliminate the communist forces remaining in the south, while northern leaders continued to strengthen their own military forces in preparation for a future confrontation. By late 1974 Hanoi had decided that victory could be achieved only through armed struggle, and early the next year North Vietnamese troops launched a major offensive against the south. Saigon's forces retreated in panic, and President Thieu ordered the abandonment of several northern provinces. Thieu's effort to stabilize the situation was too late, however, and on April 30, 1975, the communists entered Saigon in triumph.

The Socialist Republic of Vietnam. Following the communist victory, Vietnam was reunited on July 2, 1976, when the Socialist Republic of Vietnam was officially proclaimed, with its capital at Hanoi. The country faced formidable problems. In the south, millions of people had been made homeless by the war, and more than one-seventh of the population had been killed or wounded; the costs in the north were probably as high or higher. Plans to reconstruct the country called for expanding industry in the north and agriculture in the south, but it quickly became clear that these goals would be difficult to achieve.

Reunification of the country

Hanoi had been at war for more than a generation—indeed, Ho Chi Minh had died in 1969—and the bureaucracy was poorly trained to deal with the problems of peacetime economic recovery. The government encountered considerable resistance to its policies, particularly in the huge metropolis of Saigon (renamed Ho Chi Minh City in 1976), where members of the commercial sector—many of whom were ethnic Chinese—resisted the new socialist economic measures and assignment to "new economic zones" in the countryside. During the late 1970s the country also suffered major floods and drought that severely reduced food production. When the regime suddenly announced a program calling for the socialization of industry and agriculture in the south in early 1978, hundreds of thousands of people (mainly ethnic Chinese) fled the country on foot or by boat.

These internal difficulties were compounded by problems in foreign affairs. The regime pursued plans to form alliances with new revolutionary governments in neighbouring Laos and Cambodia (Kampuchea), which risked incurring not only the hostility of the United States but also that of China, which had its own interests in those countries. As Sino-Vietnamese relations soured, Hanoi signed a treaty of friendship and cooperation with the Soviet Union. Meanwhile, relations with Cambodia rapidly deteriorated when it rejected Hanoi's offer of a close relationship among the three countries of the former French Indochina. Savage border fighting culminated in a Vietnamese invasion of Cambodia in December 1978. The Khmer Rouge were dislodged from power, and a pro-Vietnamese government was installed in Phnom Penh.

Khmer Rouge forces took refuge in isolated areas of the country and began a guerrilla war of resistance against the new government, the latter backed by some 200,000 Vietnamese troops. In the meantime, China launched a brief but fierce punitive invasion along the Sino-Vietnamese border in early 1979 in response to Vietnamese actions in Cambodia. During the month-long war the Chinese destroyed major Vietnamese towns and inflicted heavy damage in the frontier zone, but they also suffered heavy casualties from the Vietnamese defenders.

Vietnam was now nearly isolated in the world. Apart from the protégé regime in Phnom Penh and the government of Laos, which also was heavily dependent on Vietnamese aid for its survival, the country was at odds with the remainder of its regional neighbours. The member states of the Association of Southeast Asian Nations opposed Vietnam's occupation of Cambodia and joined with China in supporting the Khmer Rouge and various noncommunist Cambodian groups. The United States and most other Western countries imposed a trade embargo on Vietnam. Only the Soviet Union and its eastern European allies stood by Vietnam.

Under such severe external pressure, Vietnam suffered continuing economic difficulties. The cost of stationing troops in Cambodia and of defending the border with China was heavy. To make matters worse, the regime encountered continuing problems in integrating the southern provinces into a socialist economy. In 1986 the party launched an economic reform program (*doi moi*). Implementation did not begin until 1988, however, when an economic crisis and declining Soviet support forced the party to cut spending, court foreign investment, liberalize trade, and permit free market activities.

Economic reforms

These measures stabilized the economy, but the collapse of communism in eastern Europe and disintegration of the Soviet Union left Vietnam isolated. Vietnam sought to improve relations with other countries by completely withdrawing its forces from Cambodia in September 1989 and participating in a peace conference in Paris that ended the conflict in 1991. This settlement permitted the normalization of relations with China, Japan, and Europe. Vietnam's assistance in determining the fate of Americans missing-in-action led the United States to lift the embargo in 1994 and establish diplomatic relations with Hanoi in 1995. Admission to membership in the Association of Southeast Asian Nations in July 1995 symbolized Vietnam's acceptance into the family of nations.

Peace allowed Vietnam to concentrate on the economic reforms begun in the late 1980s, and the economy grew rapidly through the 1990s. Success, however, has reduced the commitment to change and raised concern about preserving Vietnam's "socialist orientation," evident in the continued prominence of state-owned enterprises. Leaders have also worried that the corruption, inequality, and materialism associated with the market economy could undermine support for the party, and noncommunists have been excluded from power. Nonetheless, in July 2000 Vietnam signed a trade agreement with the United States that was a major step toward membership in the World Trade Organization. In that same month Vietnam opened its first stock exchange. (M.E.O./W.J.D./Ed.)

For later developments in the history of Vietnam, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

General works. GEORGE KURIAN (ed.), *Encyclopedia of the Third World*, 4th ed., 3 vol. (1992), and *Atlas of the Third World*, 2nd ed. (1992); and RICHARD ULACK and GYULA PAUER, *Atlas of Southeast Asia* (1989), contain general descriptions of the individual countries. Comprehensive annual publications include *The Far East and Australasia* and *Asia Yearbook*.

Physical and human geography. Two classic geography texts are E.H.G. DOBBY, *South East Asia*, 11th ed. (1973); and CHARLES A. FISHER, *South-East Asia*, 2nd ed. (1966). KEITH BUCHANAN,

The Southeast Asian World (1967), is older but still a useful overview. JOSEPH E. SPENCER and WILLIAM L. THOMAS, *Asia, East by South*, 2nd ed. (1971); JOSEPH E. SPENCER, *Oriental Asia* (1973); and R.D. HILL (ed.) *South-East Asia* (1979), also are valuable. Detailed studies of specific topics include CHARLES S. HUTCHISON, *Geological Evolution of South-East Asia* (1988); K. TAKAHASHI and H. ARAKAWA (eds.), *Climates of Southern and Western Asia* (1981); N. MARK COLLINS, JEFFREY A. SAYER, and TIMOTHY C. WHITMORE (eds.), *The Conservation Atlas of Tropical Forests: Asia and the Pacific* (1991); and TIMOTHY C. WHITMORE, *Tropical Rain Forests of the Far East*, 2nd ed. (1984).

Encyclopedia of World Cultures, vol. 5, *East and Southeast Asia*, ed. by PAUL HOCKINGS (1993), contains good introductory essays on the region's numerous ethnic groups; and FRANK M. LEBAR, GERALD C. HICKEY, and JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964), though dated, is still a good general work. RONALD PROVENCHER, *Mainland Southeast Asia* (1975), is an anthropological treatment. VICTOR PURCELL, *The Chinese in Southeast Asia*, 2nd ed. (1965, reissued 1980), is a classic work. GEHAN WIJEYEWARDENE (ed.), *Ethnic Groups Across National Boundaries in Mainland Southeast Asia* (1990), concentrates on the Thai border area. Also useful are GUY HUNTER, *South-East Asia—Race, Culture, and Nation* (1966); ROBBINS BURLING, *Hill Farms and Padi Fields: Life in Mainland Southeast Asia* (1965); JONATHAN RIGG, *Southeast Asia: A Region in Transition* (1991); and FRED R. VON DER MEHDEN, *Religion and Modernization in Southeast Asia* (1986). Among works on urbanization are T.G. MCGEE, *The Southeast Asian City* (1967), older but still useful; and PAUL WHEATLEY, *Nāgara and Commandery: Origins of the Southeast Asian Urban Traditions* (1983).

General surveys of the region's economy include CHRIS DIXON, *South East Asia in the World-Economy* (1991); and BRIAN WAWN, *The Economies of the ASEAN Countries* (1982). Among the studies of economic development are DONALD W. FRYER, *Emerging Southeast Asia*, 2nd ed. (1979), a well-written, insightful book; DENIS DWYER (ed.), *South East Asian Development* (1990); DAVID DRAKAKIS-SMITH, *Pacific Asia* (1992), a brief text; and RICHARD ROBISON, KEVIN HEWISON, and RICHARD HIGGOTT (eds.), *Southeast Asia in the 1980s: The Politics of Economic Crisis* (1987). Discussions of resource utilization include GEORGE KENT and MARK J. VALENCIA (eds.), *Marine Policy in Southeast Asia* (1985); LIM TECK GHEE and MARK J. VALENCIA (eds.), *Conflict Over Natural Resources in South-east Asia and the Pacific* (1990); and MARK J. VALENCIA, *South-east Asian Seas: Oil Under Troubled Waters* (1985). INTERNATIONAL ASSOCIATION OF AGRICULTURAL ECONOMISTS, *World Atlas of Agriculture*, vol. 2, *Asia and Oceania* (1973), offers a good background treatment. THOMAS R. LEINBACH and CHIA LIN SIEN, *South-East Asian Transport* (1989), discusses regional transport development. (T.R.L.)

History. *General:* A comprehensive overview is NICHOLAS TARLING (ed.), *The Cambridge History of Southeast Asia*, 2 vol. (1992); while MILTON OSBORNE, *Southeast Asia*, 5th ed. (1990), is a brief survey. JOHN FRANK CADY, *Southeast Asia* (1964), though older and marred by some factual errors, is well-organized. D.G.E. HALL, *A History of South-East Asia*, 4th ed. (1981), while thorough, is heavily slanted toward colonial topics and views.

Southeast Asia to c. 1750: PETER BELLWOOD, *Prehistory of the Indo-Malaysian Archipelago* (1985), is detailed and thought-provoking. CHARLES HIGHAM, *The Archaeology of Mainland Southeast Asia: From 10,000 B.C. to the Fall of Angkor* (1989), complements Bellwood, with more focus on archaeology. Collections of essays include DAVID G. MARR and A.C. MILNER (eds.), *Southeast Asia in the 9th to 14th Centuries* (1986); and R.B. SMITH and W. WATSON (eds.), *Early South East Asia* (1979). ANTHONY REID, *Southeast Asia in the Age of Commerce, 1450–1680*, 2 vol. (1988–93), provides a broad-based glimpse of the region that previously had not been available. DONALD F. LACH, *Southeast Asia in the Eyes of Europe: The Sixteenth Century* (1968), contains a selection of travel accounts by Europeans. ANTHONY REID and LANCE CASTLES (eds.), *Pre-colonial State Systems in Southeast Asia* (1975), includes several specific treatments. The classic by M.A.P. MEILINK-ROELOFSZ, *Asian Trade and European Influence in the Indonesian Archipelago Between 1500 and About 1630* (1962), is still useful. LORRAINE GESICK (ed.), *Centers, Symbols, and Hierarchies* (1983), is a collection of essays on the region's classical states.

Southeast Asia since c. 1750: DAVID JOEL STEINBERG *et al.*, *In Search of Southeast Asia: A Modern History*, rev. ed. (1987), is a sophisticated treatment, but its focus shifts from era to era. JOHN BASTIN and HARRY J. BENDA, *A History of Modern Southeast Asia: Colonialism, Nationalism, and Decolonization* (1968), although dated, is still worthy of careful attention. SYED HUSSEIN ALATAS, *The Myth of the Lazy Native: A Study of the Image of the Malays, Filipinos, and Javanese From the 16th to the 20th Century and Its Function in the Ideology of Colonial Capitalism* (1977), is a convincing attack by a Southeast Asian intellectual on colonialism and colonial scholarship in the region. D.J.M. TATE, *The Making of Modern South-East Asia*, 2 vol. (1971–79), treats the middle portion of the colonial age in detail. DAVID K. WYATT and ALEXANDER WOODSIDE (eds.), *Moral Order and the Question of Change* (1982), explores social and intellectual history. FRED R. VON DER MEHDEN, *South-East Asia, 1930–1970: The Legacy of Colonialism and Nationalism* (1974), although dated, is a well-illustrated, useful introduction to the postwar region. RONALD D. PALMER and THOMAS J. RECKFORD, *Building ASEAN* (1987), offers a basic introduction to the organization's first 20 years. Each essay in ALFRED W. MCCOY (ed.), *Southeast Asia Under Japanese Occupation* (1974), has its own locale and slant, although it cannot substitute for a general history of the occupation. JAN PLUVIER, *South-East Asia from Colonialism to Independence* (1974), is the most thorough treatment of the period 1942–*c.* 1965. CLARK D. NEHER, *Southeast Asia in the New International Era* (1991), has a political-science emphasis. (W.H.F.)

Brunei. Perhaps the most reliable and up-to-date reference source on the country is *Country Profile: Malaysia, Brunei* (annual). A collection of papers on the country's natural resources and environment is found in a special issue of *Singapore Journal of Tropical Geography*, vol. 13, no. 1 (June 1992). CHUA THIA-ENG, CHOU LOKE MING, and MARIE SOL M. SADORRA (eds.), *The Coastal Environmental Profile of Brunei Darussalam* (1987), includes articles on land use, population, and the institutional framework. Other economic studies include TILAK DOSHI, "Brunei: The Steady State," *Southeast Asian Affairs* (1991), pp. 71–80; and SRITUA ARIEF, *The Brunei Economy* (1986).

D.E. BROWN, *Brunei* (1970), chronicles the history of the sultanate from the early 19th century to the late 1960s. A well-researched standard text is RANJIT SINGH, *Brunei, 1839–1983* (1984). K.U. MENON, "Brunei Darussalam in 1986: In Search of the Political Kingdom," *Southeast Asian Affairs* (1987), pp. 85–101, traces political developments in the first three years of full independence. See also ABU BAKAR HAMZAH, "Brunei Darussalam Continuity and Tradition," *Southeast Asian Affairs* (1989), pp. 91–104; and ZAINAL KLING, "The Changing International Image of Brunei," *Southeast Asian Affairs* (1990), pp. 89–100. (O.J.B.)

Cambodia. General introductions are RUSSELL R. ROSS (ed.), *Cambodia: A Country Study*, 3rd ed. (1990); DAVID P. CHANDLER, *The Land and People of Cambodia* (1992); and MICHAEL VICKERY, *Kampuchea: Politics, Economics, and Society* (1986). JEAN DELVERT, *Le Paysan cambodgien* (1961), is a magisterial work. BEN KIERNAN and CHANTHOU BOUA (eds.), *Peasants and Politics in Kampuchea, 1942–1981* (1982), is a useful anthology. RÉMY PRUD'HOMME, *L'Économie du Cambodge* (1969), is the only detailed study of the Cambodian economy and is still of historical interest.

DAVID P. CHANDLER, *A History of Cambodia*, 2nd ed. (1992), traces the country's beginnings through the 1980s and is supplemented by his *The Tragedy of Cambodian History* (1991), a detailed political history since World War II. CLAUDE JACQUES, *Angkor* (1990), is an up-to-date treatment (in French) by France's leading Angkorean scholar. An indispensable work for scholars is LAWRENCE PALMER BRIGGS, *The Ancient Khmer Empire* (1951, reprinted 1962). In addition to the works by Chandler above, modern Cambodian history is analyzed in BEN KIERNAN, *How Pol Pot Came to Power* (1983); WILLIAM SHAWCROSS, *Sideshow: Kissinger, Nixon, and the Destruction of Cambodia*, rev. ed. (1987); KARL D. JACKSON (compiler), *Cambodia, 1975–1978* (1989); and MICHAEL VICKERY, *Cambodia, 1975–1982* (1984), probably the best book-length analysis of the revolutionary era. (D.P.Ch.)

Laos. FRANK M. LEBAR and ADRIENNE SUDDARD (eds.), *Laos*, rev. ed. (1967), is a general survey. GRANT EVANS, *Lao Peasants Under Socialism* (1990), treats ethnography and economics. MARTIN STUART-FOX, *Laos* (1986), is the standard work, with good chapters on the economic system and domestic policies. A more current overview is found in WILLIAM WORNER, "Economic Reform and Structural Change in Laos," *Southeast Asian Affairs* (1989), pp. 187–208. JOSEPH J. ZASLOFF and LEONARD UNGER (eds.), *Laos: Beyond the Revolution* (1991), examines political, economic, social, and foreign policies. Further information can be found in the annotated bibliography by HELEN CORDELL (compiler), *Laos* (1991). (A.J.D.)

ARTHUR J. DOMMEN, *Laos: Keystone of Indochina* (1985), is a brief general history, and his *Conflict in Laos*, rev. ed. (1971), a political history, focuses primarily on the period from the early 1950s to 1970. HUGH TOYE, *Laos* (1968), depicts Laos' historic position between Vietnam and Thailand. PAUL F. LANGER and JOSEPH J. ZASLOFF, *North Vietnam and the Pathet Lao* (1970), studies the role of communist North Vietnam in the origin and development of the Lao communist movement. JOSEPH J. ZASLOFF, *The Pathet Lao* (1973), examines the political dynamics

of the group, its leadership, commanding party, front, political and administrative organizations, and military forces. MARTIN STUART-FOX (ed.), *Contemporary Laos* (1982), is a collection of essays. MACALISTER BROWN and JOSEPH J. ZASLOFF, *Apprentice Revolutionaries: The Communist Movement in Laos, 1930–1985* (1985), provides a political analysis. (J.J.Z.)

Malaysia. R.S. MILNE and DIANE K. MAUZY, *Malaysia* (1986), is a comprehensive overview. OOI JIN-BEE, *Peninsular Malaysia*, new ed. (1976), offers a good geographic overview of the country. JAMES C. JACKSON, *Sarawak* (1968), is one of the few quality studies of the state. Studies of Malaysia's people include JUDITH NAGATA, *Malaysian Mosaic: Perspectives from a Polyethnic Society* (1979); KERNIAL SINGH SANDHU, *Indians in Malaya: Some Aspects of Their Immigration and Settlement (1786–1957)* (1969); HEATHER STRANGE, *Rural Malay Women in Tradition and Transition* (1981); HENG PEK KOON, *Chinese Politics in Malaysia: A History of the Malaysian Chinese Association* (1988); and JAMES V. JESUDASON, *Ethnicity and the Economy: The State, Chinese Business, and Multinationals in Malaysia* (1989). RAJ KUMAR, *The Forest Resources of Malaysia, Their Economics and Development* (1986); and S. ROBERT AIKEN *et al., Development and Environment in Peninsular Malaysia* (1982), focus on both environmental concerns and economic development. Other economic studies include MOHAMED ARIFF, *The Malaysian Economy* (1991); and GEORGE CHO, *The Malaysian Economy: Spatial Perspectives* (1990). See also E.K. FISK and H. OSMAN-RANI (eds.), *The Political Economy of Malaysia* (1982); and GORDON P. MEANS, *Malaysian Politics: The Second Generation* (1991). (T.R.L.)

The best comprehensive history of Malaysia is BARBARA WATSON ANDAYA and LEONARD Y. ANDAYA, *A History of Malaysia* (1982). A good survey is JOHN GULLICK, *Malaysia: Economic Expansion and National Unity* (1981). Both of these works are stronger on Malaya than on the Borneo states. STEVEN RUNCIMAN, *The White Rajahs: A History of Sarawak from 1841 to 1946* (1960); and ROBERT PRINGLE, *Rajahs and Rebels: The Ibans of Sarawak Under Brooke Rule, 1841–1941* (1970), provide excellent coverage of the Brooke era. Among the few detailed accounts of ancient Malaya, PAUL WHEATLEY, *The Golden Khersonese: Studies in the Historical Geography of the Malay Peninsula Before A.D. 1500* (1961, reprinted 1973), remains a classic work. LEONARD Y. ANDAYA, *The Kingdom of Johor, 1641–1728* (1975), is a fine analysis of that sultanate. LIM TECK GHEE, *Peasants and Their Agricultural Economy in Colonial Malaya, 1874–1941* (1977), treats economic history during the colonial era. WILLIAM R. ROFF, *The Origins of Malay Nationalism* (1967), is a stimulating work that explores Malay society in the colonial years. VICTOR PURCELL, *The Chinese in Malaya* (1948, reissued 1967), though somewhat dated, remains the only general survey. R.S. MILNE and DIANE K. MAUZY, *Politics and Government in Malaysia*, rev. ed. (1980), is particularly good on the late colonial and early independence periods. (C.A.Lo.)

Myanmar. FREDERICA M. BUNGE (ed.), *Burma: A Country Study*, 3rd ed. (1983), is an overview. Many of the English-language works on Myanmar's geography were produced by the British during the colonial period. Among these are the still invaluable *Burma Gazetteer*, 30 vol. (1868–1935), with detailed surveys of different administrative districts; *The British Burma Gazetteer*, 2 vol. (1879–80, reprinted as *Gazetteer of Burma*, 1987); J. GEORGE SCOTT and J.P. HARDIMAN (compilers), *Gazetteer of Upper Burma and the Shan States*, 3 vol. in 5 (1900–01, reprinted 2 vol. in 5, 1983); and H.L. CHHIBBER, *The Physiography of Burma* (1933, reprinted 1975). Also dated but still useful is HELLMUT DE TERRA and HALLAM L. MOVIUS, JR., "Research on Early Man in Burma," *Transactions of the American Philosophical Society*, new series, 32(3):267–393 (1943). MICHAEL AUNG-THWIN, *Irrigation in the Heartland of Burma* (1990), studies the productive capacity and geography of pre-colonial Myanmar. Agriculture is detailed in M.Y. NUTTONSON, *The Physical Environment and Agriculture of Burma* (1963), a brief, technical study; CHENG SIOK-HWA, *The Rice Industry of Burma, 1852–1940* (1968), the only book-length study in English covering the topic during that period; and U KHIN WIN, *A Century of Rice Improvement in Burma* (1991), a more recent study.

MICHAEL AUNG-THWIN, "Spirals in Early Southeast Asian and Burmese History," *The Journal of Interdisciplinary History*, 21(4):575–602 (Spring 1991), is a theoretical and conceptual treatment of broad historical patterns. G.H. LUCE, *Phases of Pre-Pagán Burma*, 2 vol. (1985), studies the 9th century in detail. AUNG THAW, *Historical Sites in Burma* (1972), deals with the pre-Pagan period. G.H. LUCE *et al., Old Burma—Early Pagán*, 3 vol. (1969–70), is the classic study of this kingdom. MICHAEL AUNG-THWIN, *Pagan: The Origins of Modern Burma* (1985), studies the kingdom's institutional history. PAUL J. BENNETT, "The 'Fall of Pagan': Continuity and Change in 14th-Century Burma," in his *Conference Under the Tamarind Tree: Three Essays in Burmese History* (1971), pp. 3–53, assesses the manner in which continuity and change made their mark on early Myanmar history. MICHAEL AUNG-THWIN, "The Role of Sasana Reform in Burmese History: Economic Dimensions of a Religious Purification," *The Journal of Asian Studies*, 38(4):671–688 (August 1979), details the economic relationship between church and state. VICTOR B. LIEBERMAN, *Burmese Administrative Cycles: Anarchy and Conquest, c. 1580–1760* (1984), analyzes the Toungoo dynasty. WILLIAM J. KOENIG, *The Burmese Polity, 1752–1819* (1990), studies the early period of the last Myanmar dynasty. OLIVER B. POLLAK, *Empires in Collision: Anglo-Burmese Relations in the Mid-Nineteenth Century* (1979), treats British policy and its effects on later colonization. DOROTHY WOODMAN, *The Making of Burma* (1962), is the most thorough account of theretofore secret British decisions in the colonization of Burma. MICHAEL ADAS, *The Burma Delta: Economic Development and Social Change on an Asian Rice Frontier, 1852–1941* (1974), traces the agricultural development of this area and its significance on modern history. JOHN F. CADY, *A History of Modern Burma* (1958, reissued 1965), is still the most complete text on the early years of the modern country. JOSEF SILVERSTEIN, *Burmese Politics* (1980), offers a Western perspective on 20th-century politics. More recent studies on the modern state include ROBERT H. TAYLOR *The State in Burma* (1987); DAVID I. STEINBERG, *Burma: A Socialist Nation of Southeast Asia* (1982); and MARTIN SMITH, *Burma: Insurgency and the Politics of Ethnicity* (1991). (M.A.A.-T.)

Singapore. Overviews are provided by PHILIPPE REGNIER, *Singapore: A City-State in South East Asia* (1991); and R.S. MILNE and DIANE K. MAUZY, *Singapore: The Legacy of Lee Kuan Yew* (1990). TANIA LI, *Malays in Singapore* (1989), is a thorough study. Economic development and government policies are surveyed in PETER S.J. CHEN, *Singapore Development Policies and Trends* (1983); LIM CHONG-YAH and PETER J. LLOYD (eds.), *Singapore: Resources and Growth* (1986); LINDA LIM and PANG ENG FONG, *Trade, Employment, and Industrialisation in Singapore* (1986), an excellent analysis; TILAK DOSHI, *Houston of Asia: The Singapore Petroleum Industry* (1989); GARRY RODAN, *The Political Economy of Singapore's Industrialization* (1989), a discussion of the role of the government in engineering development; FREDERICK DEYO, "Singapore: Developmental Paternalism," in STEVEN M. GOLDSTEIN (ed.), *Minidragons: Fragile Economic Miracles in the Pacific* (1991), pp. 48–87, a candid article on the impact of state policies; and KERNIAL SINGH SANDHU and PAUL WHEATLEY (eds.), *Management of Success: The Moulding of Modern Singapore* (1989), a collection of essays.

C.M. TURNBULL, *A History of Singapore, 1819–1988*, 2nd ed. (1989), is the key historical study on the nation, with an excellent bibliography. DONALD MOORE and JOANNA MOORE, *The First 150 Years of Singapore* (1969), is an older but still useful work. A more recent survey is ERNEST C.T. CHEW and EDWIN LEE (eds.), *A History of Singapore* (1991). YEN CHING-HWANG, *A Social History of the Chinese in Singapore and Malaya, 1800–1911* (1986), is written from Chinese records and accounts. L.K. WONG, "Singapore: Its Growth as an Entrepot Port, 1819–1941," *Journal of Southeast Asian Studies*, 9(1):50–84 (March 1978), is the best treatment of economic history up to World War II. ALEX JOSEY, *Singapore* (1979), stresses the contemporary period from a pro-People's Action Party stance. (T.R.L.)

Thailand. Surveys of the country include ELLIOTT KULICK and DICK WILSON, *Thailand's Turn: Profile of a New Dragon* (1992); and FRANK J. MOORE and CLARK D. NEHER, *Thailand—Its People, Its Society, Its Culture* (1974). ROBERT L. PENDLETON, *Thailand* (1962, reprinted 1976), still provides a clearly presented and readable description of physical geography. PAUL LEWIS and ELAINE LEWIS, *Peoples of the Golden Triangle* (1984), describes six northern tribes; while SERI PHONGPHIT and KEVIN HEWISON, *Thai Village Life* (1990), contains an excellent portrayal of the people of the northeast. WOLF DONNER, *The Five Faces of Thailand* (1978), surveys economic geography in detail through the 1970s. LUCIEN M. HANKS, *Rice and Men: Agricultural Ecology in Southeast Asia* (1972, reissued 1992), although dated, is one of the best introductions to daily life and social and economic changes in rice-growing villages in central Thailand. WILLIAM J. KLAUSNER, *Reflections on Thai Culture*, 3rd ed. (1987), collects articles on village life in the northeast and on Thai culture and values. (Ja.A.H.)

DAVID K. WYATT, *Thailand* (1984), is a major reference work giving a detailed historical overview. Other general studies include CHARLES KEYES, *Thailand: Buddhist Kingdom as Modern Nation-State* (1987). CHARNVIT KASETSIRI, *The Rise of Ayudhya: A History of Siam in the Fourteenth and Fifteenth Centuries* (1976), is a detailed survey. AKIN RABIBHADANA, *The Organization of Thai Society in the Early Bangkok Period, 1782–1873* (1969), is a basic source on traditional 19th-century society. HONG LYSA, *Thailand in the Nineteenth Century* (1984), is the most comprehensive socioeconomic history of this period.

WALTER F. VELLA and DOROTHY B. VELLA, *Chaiyo!: King Vajiravudh and the Development of Thai Nationalism* (1978), is a thorough study of the early 20th century. BENJAMIN A. BATSON, *The End of the Absolute Monarchy in Siam* (1984), analyzes the political and social conditions of the 1920s and '30s that set the stage for later events. JUDITH A. STOWE, *Siam Becomes Thailand* (1991), describes the period from the abolition of the absolute monarchy in 1932 through World War II. THAK CHALOEMTIARANA, *Thailand: The Politics of Despotic Paternalism* (1979), analyzes the rise of the military from the return of Phibun in 1947 to the Sarit coup of 1957 and discusses the character of military rule in Thailand. JOHN L.S. GIRLING, *Thailand* (1981), focuses especially on the period 1963–77; and DAVID MORELL and CHAI-ANAN SAMUDAVANIJA, *Political Conflict in Thailand* (1981), is a detailed study of the 1970s.

(E.J.K./C.F.Ke.)

Vietnam. DAVID W.P. ELLIOT *et al., Vietnam: Essays on History, Culture, and Society* (1985), is a solid and readable introduction. A comprehensive analysis of Vietnamese society and culture is NEIL L. JAMIESON, *Understanding Vietnam* (1993). Other studies of peoples and society include PIERRE GOUROU, *The Peasants of the Tonkin Delta,* 2 vol. (1955, originally published in French, 1936), a monumental work; three books by GERALD C. HICKEY, *Village in Vietnam* (1964), a classic ethnography of the upper Mekong delta in the 1950s, *Sons of the Mountains* (1982), a masterful overview of the evolving cultures of the Montagnards up to 1954, and *Free in the Forest* (1982), a scholarly description of the fate of the peoples of the central highlands from 1954 to 1976; and two studies by A. TERRY RAMBO, *A Comparison of Peasant Social Systems of Northern and Southern Viet-Nam* (1973), and "Vietnam: Searching for Integration," in CARLOS CALDAROLA (ed.), *Religions and Societies, Asia and the Middle East* (1982), pp. 407–444. NIGEL THRIFT and DEAN FORBES, *The Price of War: Urbanization in Vietnam, 1954–85* (1986); VO NHAN TRI, *Vietnam's Economic Policy Since 1975* (1990); and MELANIE BERESFORD, *Vietnam: Politics, Economics, and Society* (1988), are useful studies on these topics. (N.L.J.)

KEITH WELLER TAYLOR, *The Birth of Vietnam* (1983), is the definitive treatment of early history to the 10th century. JOSEPH BUTTINGER, *The Smaller Dragon: A Political History of Vietnam* (1958, reissued 1966), is the standard history from the rise of the Vietnamese state to the colonial era. THOMAS HODGKIN, *Vietnam* (1981), recounts in detail the background of the Vietnamese revolutionary struggle. ALEXANDER B. WOODSIDE, *Vietnam and the Chinese Model* (1971, reprinted 1988), analyzes 19th-century Vietnamese society. DAVID G. MARR, *Vietnamese Anticolonialism, 1885–1925* (1971), is a sensitive account, while his *Vietnamese Tradition on Trial, 1920–1945* (1981), explores the social and intellectual changes taking place under colonial rule. MILTON E. OSBORNE, *The French Presence in Cochinchina and Cambodia* (1969), analyzes French policies during the first stages of colonial rule. ALEXANDER B. WOODSIDE, *Community and Revolution in Modern Vietnam* (1976), argues that the search for community is a key factor in the Vietnamese revolution. HUYNH KIM KHANH, *Vietnamese Communism, 1925–1945* (1982), is the definitive account of the rise of the Vietnamese communist movement. ELLEN HAMMER, *The Struggle for Indochina* (1954, reissued 1969), dramatically treats the final stages of French rule. WILLIAM J. DUIKER, *The Communist Road to Power in Vietnam* (1981), is a historical study of the evolution of Vietnamese communist strategy. BERNARD B. FALL, *The Two Viet-Nams,* 2nd rev. ed. (1967), dated but still useful, examines the period after the division of the country. STANLEY KARNOW, *Vietnam: A History,* rev. and updated ed. (1991), is the companion volume to the PBS film series on the Vietnam War. DOUGLAS PIKE, *Viet Cong: The Organization and Techniques of the National Liberation Front of South Vietnam* (1966), is the classic analysis of communist techniques during the war; while TRUONG NHU TANG, DAVID CHANOFF, and DOAN VAN TOAI, *A Viet Cong Memoir* (1985, also published as *Journal of a Viet Cong,* 1986), is a firsthand account of the organization. KEN POST, *Revolution, Socialism, and Nationalism in Viet Nam,* 4 vol. (1989–92), is a Marxist interpretation of the Vietnamese revolution. JEFFREY RACE, *War Comes to Long An* (1972), discusses the factors behind the communist insurgency in South Vietnam, viewed from a single province. WILLIAM J. DUIKER, *Vietnam Since the Fall of Saigon,* updated ed. (1989), chronicles Hanoi's domestic and foreign policies since the end of the war. ROBERT SHAPLEN, *Bitter Victory* (1986), provides a journalistic account of Hanoi's problems in winning the peace since the end of the war. (W.J.D.)

Southeast Asian Arts

The term Southeast Asia refers to the huge peninsula of Indochina and the extensive archipelago of what is sometimes called the East Indies. The region can be subdivided into mainland Southeast Asia and insular Southeast Asia. The political units contained in this region are Burma, Thailand, Laos, Kampuchea (Cambodia), Vietnam, Malaysia, Singapore, Indonesia, and the Philippines. The Philippines originally was not included, because Philippine history has not followed the general historical pattern of Southeast Asia, but, because of its geographic position and the close affinities of its primitive cultures with the primitive cultures of Southeast Asia, it is now usually regarded as the eastern fringe of Southeast Asia. A common geographic and climatic pattern prevails over all of Southeast Asia and has resulted in a particular pattern of settlement and cultural development. Mountain people generally have a cultural level less developed than that of the valley dwellers. As a consequence, Southeast Asia is culturally fragmented. (For a discussion of the traditional cultures of the area, see ASIAN PEOPLES AND CULTURES: *Southeast Asian cultures.*)

The article is divided into the following sections:

The cultural setting of Southeast Asian arts 795
External influences
Indigenous traditions
The unique aesthetic of the region
Literature 797
General considerations
Pre-European colonial period
European colonial and modern periods
Music 800
General characteristics
Historical developments
The performing arts 805
Diverse traditions in the performing arts
Characteristics of dance
Characteristics of drama
Origins and development of the performing arts
Diverse national forms and traditions
Visual arts 813
General considerations
Burma
Thailand and Laos
Cambodia and Vietnam
Indonesia
The Philippines
Folk arts
Bibliography 836

The cultural setting of Southeast Asian arts

Southeast Asia has been the crossroads of many races who have been contending against each other for centuries. At present, all the peoples of Southeast Asia are Mongoloid in racial origin. The first to come were the Austronesians (Malayo-Polynesians), sometimes described as Proto-Malays and Deutero-Malays. At one time they occupied the eastern half of mainland Southeast Asia, but later they were pushed toward the south and the islands by the Austro-Asiatics. At present, peoples of Austronesian origin occupy Malaysia, the Republic of Indonesia, and the Republic of the Philippines. There were three main Austro-Asiatic races, the Mons, the Khmers, and the Viet-Muong. The Mons were at one time dominant, but they lost their racial identity in the 18th century and became absorbed by the Burmese and the Tais; only a few thousand Mons are now found living near the Burma–Thailand border. The Khmers from the 9th century to the 15th built a great empire, but much of its territory was lost to its neighbours so that only the small kingdom of Cambodia remains today. The Viet-Muong now occupy Vietnam. A Tibeto-Burmese tribe, the Pyu, founded an empire of city-kingdoms in the Irrawaddy Valley in the early centuries of the Christian Era, but the Pyu disappeared, and the Burmese, taking the leadership, founded their kingdom of Pagan and have occupied Burma up to the present day. In the 13th century the Tai-Shan lost their kingdom of Nan-chao in Yunnan, China, and entered the Mae Nam Chao Phraya Valley to found kingdoms that gradually evolved into the kingdoms of Siam (Thailand) and Laos.

EXTERNAL INFLUENCES

The Indian influence

In Southeast Asia, winds of change often came as storms. Indian commerce expanded into Southeast Asia in the early centuries of the Christian Era and, in spite of its peaceful nature, caused revolutionary changes in the life and culture of the peoples of the region. The Indians would sojourn in the region in small numbers for two or three monsoons only. The success of their commercial venture and the safety of their persons depended entirely on the goodwill of the inhabitants. At that time, the Indians were at a higher level of culture, and they brought new ideas and new art traditions. Since these ideas had some affinity with indigenous ideas and art forms, the natives, who had already reached a high level of culture themselves, accepted them but were not overwhelmed by an influx of new traditions. The Hindu-Buddhist culture of the Indians made a tremendous impact and came to form the second layer of culture in Southeast Asia, but the first layer of native ideas and traditions has remained strong to the present day.

European influence

Changes often came to Southeast Asia, usually because it possessed a commodity that was in great demand by the rest of the world. The Indians came because they were looking for fresh sources of gold after the Roman imperial source had run dry. In the 15th, 16th, and the 17th centuries, insular Southeast Asia attracted Islāmic merchants from India and farther west and later the Portuguese and the Dutch as a rich source of spices. As with the Hindu–Buddhist merchants of the past, the Islāmic traders came not as missionaries, though they did spread their religion in the region. The Portuguese came as conquerors and as militant missionaries of their Roman Catholic form of Christianity, and, for those reasons, their cultural traditions were found unacceptable to the natives. In the 17th century the Dutch came as conquerors and colonists for whom the attraction was first spices and then coffee, rubber, and petroleum. Since mainland Southeast Asia produced no spices for export, it was less vulnerable to the navies of Portugal and the Netherlands, so the region was not greatly affected by the Muslims, Portuguese, and Dutch. In the 19th century, Britain and France became interested in mainland Southeast Asia as the back door to China and sought to possess it as a colony. By the end of the 19th century, Burma had fallen to Britain, Siam was allowed to retain its independence only with the tacit permission of the two powers, and the rest had fallen to France. When in the mid-20th century the whole of Southeast Asia became free again, it was found that the arts of the peoples had remained merely static during the years of colonial rule; in some regions they had been driven underground by long neglect, but European culture and European art forms had made only a small impact.

INDIGENOUS TRADITIONS

The peoples of Southeast Asia were once thought to have shared a lack of inventiveness since prehistoric times and to have been "receptive" rather than "creative" in their contacts with foreign civilizations. Later excavations and discoveries in Burma and Thailand, however, inspired some scholars to argue against the accepted theory that

civilization came to Southeast Asia from China in prehistoric times; rather, these scholars contend, the peoples of mainland Southeast Asia were cultivating plants, making pottery, and working in bronze about the same time as the peoples of the ancient Middle East, and therefore civilization spread from mainland Southeast Asia to China and India. Southeast Asians have never produced any theory of art or literary or dramatic criticism, for they are always more concerned with doing the actual work of producing beautiful things. Because the Southeast Asians, especially in the western half of the mainland, worked on nondurable materials, it is not possible to trace the development and evolution of art forms stage by stage. The region has always been thickly forested, so it was natural that the first material to be used for artistic purposes should have been wood. The wood-carving tradition, begun in primitive times, was retained even when they learned to work with metals and with stone and continued to flourish long after the great age of stone sculpture and stone architecture, which ended in the 13th century. Proto-Neolithic paintings discovered in a cave near the Salween River in the western Shan state of Burma have very close affinity with the later carvings on posts of houses among the Nāgas on the western hills of Burma. Similarly, cave paintings of a pair of human hands with open palms, one holding the sun and the other holding a human skull, are reflected in the later aesthetic tradition of Southeast Asia: the sun symbol is found as an art motif all over the region, and a suggestion of awe, triumph, and joy at acquiring a human head is found in carvings under the eaves of the Nāga houses. The cave painting testifies to the continuity of the magico-religious tradition connected with all the arts of the area.

The art of casting the bronze drums found at Dong Son, near Hanoi, which are similar to the bronze drums used by mountain tribes throughout Southeast Asia, was thought to have come from China, but recent excavations in Thailand proved that the drums and the so-called Dong Son culture itself are native to mainland Southeast Asia. In any case, the continuity of the aesthetic tradition of Southeast Asia can be seen in the bronze drums that were cast by the Karens of Burma for centuries until the early years of the 20th century. The mountains of mainland Southeast Asia provided gold, silver, and other metals, and the art of metalworking must have developed quite early. Silver buttons, belts, and ornaments now made and worn by the hill peoples in Southeast Asia have behind them a very ancient tradition of workmanship. The same artistic tradition is found in textile designs.

Music, dance, and song were originally associated with tribal rituals. From the beginning, the main characteristic of Southeast Asian music and dance has been a swift rhythm. The slow and stately dances of the Siamese court were of Indian origin; when they were introduced into Burma in the 16th century, the Burmese quickened the tempo, but, even with that modification, the dances were still called Siamese dances to distinguish them from the native ones. In their oral literature—namely, in folk songs and folktales—the emphasis is on gaiety and humour. Typically, Southeast Asians do not like an unhappy ending.

The role of royal patronage and religious institutions. In all the regions of Southeast Asia, the arts flourished under the patronage of the kings. About the time of the birth of Christ, tribal groups gradually organized themselves, after some years of settled life as rice cultivators, into city-kingdoms, or conglomerations of villages. A king was thus little more than a paramount tribal chieftain. Since the tribes had been accustomed to worshipping local spirits, the kings sought a new spirit that would be worshipped by the whole community. One reason why the gods of Hinduism and Buddhism were found so readily acceptable to Southeast Asia was this need for new national gods. The propagation of the new religions was the task of the kings, and consequently the period from the 1st to the 13th century was a great age of temple building all over Southeast Asia. Architecture, sculpture, and painting on the temple walls were the arts that flourished. In the ancient empires of eastern Indochina and the islands, scholars of Sanskrit, the language of the sacred works of Hinduism, became part of the king's court, producing a local Sanskrit literature of their own. This literary activity was confined to the hereditary nobility and never reached the people, except in stories from the great Hindu epics *Mahābhārata* and *Rāmāyaṇa*. Because the Hindu religious writings in Sanskrit were beyond the reach of the common people, Hinduism had to be explained to them by Hindu stories of gods and demons and mighty men. On the other side of the peninsula, in the Pyu-Burmese empire of Prome, which flourished before the 8th century, there was no such development—first, because Hinduism was never widely accepted in Burma and, second, because the more open Burmese society developed neither the institution of a god-king nor that of a hereditary nobility. Although Pāli scholars surrounded the king in later Pagan, Pāli studies were pursued not at the court but at monasteries throughout the kingdom so that even the humblest villager had some faint contact with Pāli teachings. While the courts of the kings in Cambodia and Java remained merely local centres of Sanskrit scholarship, Pagan became a centre of Pāli learning for Buddhist monks and scholars even from other lands. As in the case of stories from the Indian epics, stories of the *Jātakas* (birth stories of the Buddha) were used to explain Buddhism to the common people, who could not read the scriptures written in Pāli. Just as scenes from the great epics in carving or in fresco adorned the temples in Cambodia and Java, scenes from the *Jātakas* adorned the Pagan temples.

Sanskrit and Pāli literature

Musicians of the Pyu kingdom played before the Emperor of China in 801, and the various musical instruments at the performance have their counterparts at the present day, not only in Burma but throughout Southeast Asia. At Pagan the people were so fond of music that even the collection of taxes became an occasion to dance and sing, and a royal official, endowing a temple, inscribed a prayer asking that in all his future existences until he reached Nirvāṇa "might he be woken up every morning to the strains of music sweetly played on flute and violin." In spite of this love for music and dance, no dramatic art seems to have developed in Burma, perhaps because Sanskrit, in which there was a dramatic tradition, was not studied. In contrast, at the courts of Cambodia and Java, the Sanskrit drama, Hindu dances, and native dance traditions combined and produced the court opera ballets. These dramatic elements later reached the common people by way of the shadow play.

The patronage of the king and the religious enthusiasm of the common people could not have produced the great temples without the enormous wealth that suddenly became available in the region following the commercial expansion. With the Khmer and Javanese empires, the wealth was produced by a feudalistic society, and so the temples were built by the riches of the king and his nobles, combined with the compulsory labour of their peasants and slaves, who probably derived some aesthetic pleasure from their work because of their religious fervour. Nonetheless, their monuments, such as Borobuḍur, in Java, and Angkor Wat, in Cambodia, had an atmosphere of massive, all-conquering power. At Pagan, where wealth was shared by the king, the royal officials, and the common people, the temples and the monasteries were built by all who had enough not only to pay the artisans their wages but also to guarantee their good health, comfort, and safety during the actual construction. The temples were dedicated for use by all monks and lay people as places of worship, meditation, and study, and the kings of Pagan did not build a single tomb for themselves. The Khmer temple of Angkor Wat and the Indonesian temple of Borobuḍur were tombs in that the ashes of the builders would be enshrined therein; the kings left stone statues representing them as gods for posterity to worship, whereas at Pagan there was only one statue of a king, and it depicted him on his knees with his hands raised in supplication to the Buddha. Consequently, the atmosphere that pervaded the temples of Pagan was one of joy and tranquillity.

Significance of the temples

This golden age of wealth and splendour in Southeast Asia ended in the 13th century with a sudden violence, when Kublai Khan's Tatar Chinese armies destroyed both the Burmese and the Khmer empires and his navy attacked Vietnam and Java. The tiny kingdoms that subsequently

sprang up all over Southeast Asia continually fought among themselves; their kings were neither powerful nor rich, and the royal courts became centres of military planning and political intrigue. During the 13th and 14th centuries, in the new Javanese kingdom of Majapahit and the new Burmese kingdom of Ava, vernacular literatures came into being. Again, differences in social structure had aesthetic repercussions. In Majapahit the king was powerful and gave his patronage to the newly arisen literature, confining it to the court. At Ava the vernacular literature bloomed throughout the kingdom, and the king, lacking power and prestige, prevailed upon some established writers to join the court circles and give them glamour.

After Majapahit, a new cultural force—namely, Islām—reached insular Southeast Asia, and over the two layers of primitive and Hindu-Buddhist cultures was added the third layer of Islām. In mainland Southeast Asia, a new Burmese empire arose over the ruins of the old and continued its task of spreading Buddhism. Hindu tradition reached the Burmese court secondhand in the 18th century as the result of the Burmese conquest of Siam and was one of the factors that contributed to the rise of a Burmese drama. On the other side of the peninsula, Vietnam, reconquered by China, fell more and more under the influence of Chinese culture. After a short period of Islāmic bloom, native culture in insular Southeast Asia was subjected to alien rule. In Burma and Siam alone among the states of Southeast Asia, native arts continued to flourish because, after centuries of warfare, they finally emerged as strong kingdoms.

Predominant artistic themes. The predominant themes of Southeast Asian arts have been religion and national history. In religion the main interest was not so much in actual doctrine but in the life and personality of the Buddha and the personalities and lives of the Hindu gods. In national history the interest was in the legendary heroes of the past, and this theme appeared only after the great empires had fallen and the memories of their glory and power remained. The Buddha image, which went through various stages of development, remained the favourite motif of sculpture and painting. The depiction of scenes from his previous lives in fresco and relief sculpture also had the purpose of teaching the Buddhist ethics to the people, as the *Jātaka*s emphasized certain moral virtues of the Buddha in his previous lives; it also gave an opportunity to the artist to introduce local colour by using, as background, scenes from his own contemporary time. The depiction of scenes from the Hindu epics also had the same purpose and gave the same opportunity to the artist. Many figures from the Buddhist and Hindu scriptures, such as gods and goddesses, heroes and princesses, hermits and magicians, demons and dragons, flying horses and winged maidens, became fused with similar native figures, and, gradually, folklore plots became merged in the general religious themes.

The *nāgas*

The *nāga*, a superhuman spirit, was taken from Buddhist and Hindu texts and merged with native counterparts, with the result that different images of the *nāga* appeared in various regions. The Burmese *nāga* was a snake with a crested head. The Mon *nāga* was a crocodile, and the Khmer and Indonesian *nāga* was conceived as a nine-headed snake. The demons of various kinds from all over Southeast Asia became merged under one name of Pāli-Sanskrit origin, *yakkha* or *yakṣa*, but they retained their separate identities in sculpture and paintings of their own different countries. The lion, which was unknown to the monsoon forest but was a figure of Hindu and Buddhist mythology, evolved into a native symbol and art motif. The primitive worship of the snake-dragon as a god of fertility was retained in the Khmer Empire; the nine-headed *nāga* became a symbol of security and of royalty, and stone *nāga*s guarded the palaces and temples. Buddhism frowned upon *nāga* worship. In Burmese and Mon sculpture the *nāga* was always shown as a servant of the Buddha, putting his body in coils to make a seat for his master and raising his great hood as an umbrella over his master's head. According to tradition, the guardian figure of a Mon temple was a two-bodied lion with a man's head, and the guardian figure of a Burmese temple was the crested lion. The Tais made themselves heirs to both the Khmer and the Mon art traditions relating to the *nāga*, but the guardian figure of their temples was the benevolent demon.

Primitive symbols and animal imagery merged with Indian animals and entered the arts. The Pyus embossed the primitive symbol of the sun on their coins as insignia of their power, and the Burmese transformed it into their favourite bird, the peacock, on the excuse that Buddhist mythology associated the peacock with the sun; the Mons adopted the red sheldrake as their symbol, and in Indonesia the mythical bird called Garuḍa, the vehicle of Vishnu, became merged with the local eagle. The figures of these birds also became decorative motifs. Animals of the Southeast Asian forests whose figures had adorned primitive dwellings of wood and thatch were stylized and came to adorn palaces and monasteries. Primitive geometrical patterns mixed with new spirals and curves from India, and Indian floral designs merged with those of trees and fruits and flowers copied from the monsoon forests.

THE UNIQUE AESTHETIC OF THE REGION

The arts of Southeast Asia have no affinity with the arts of other areas, except India. Burma was always an important route to China, but Burmese arts showed very little Chinese influence. The Tais, coming late into Southeast Asia, brought with them some Chinese artistic traditions, but they soon shed them in favour of the Khmer and Mon traditions, and the only indications of their earlier contact with Chinese arts were in the style of their temples, especially the tapering roof, and in their lacquer ware. Vietnam was a province of China for 1,000 years, and its arts were Chinese. The Hindu archaeological remains in southern Vietnam belong to the ancient kingdom of Champa, which Vietnam conquered in the 15th century. The Buddhist statues in northern Vietnam were Chinese Buddhist in style. The essential differences in aesthetic aim and style between the arts of East Asia and those of Southeast Asia could be seen in the contrast between the emperors' tombs of Vietnam and the temple-tombs of Cambodia and Indonesia or the opulent and dignified Buddha images of Vietnam and the ascetic and graceful Buddha images of Cambodia and Burma. Islāmic art, with its rejection of animal and human figures and its striving to express the reality behind the false beauty of the mundane world, also has no affinity with Southeast Asian arts. Both Hinduism and Buddhism taught that the sensual world was false and transitory, but this message found no place in the arts of Southeast Asia. The world depicted in Southeast Asian arts was a mixture of realism and fantasy, and the all-pervading atmosphere was a joyous acceptance of life. It has been pointed out that Khmer and Indonesian classical arts were concerned with depicting the life of the gods, but to the Southeast Asian mind the life of the gods was the life of the peoples themselves—joyous, earthy, yet divine. The European theory of "art for art's sake" found no echo in Southeast Asian arts, nor did the European division into secular and religious arts. The figures tattooed on a Burmese man's thigh were the same figures that adorned a great temple and decorated a lacquer tray. Unlike the European artist, the Southeast Asian did not need models, for he did not strive to be realistic and correct in every anatomical detail. This intrusion of fantasy and this insistence on the joyousness of human life have made Southeast Asian arts unique.

Literature

GENERAL CONSIDERATIONS

Regional distinctions. From the point of view of its "classical" literatures, Southeast Asia can be divided into three major regions: (1) the Sanskrit region of Cambodia and Indonesia; (2) the region of Burma where Pāli, a dialect related to Sanskrit, was used as a literary and religious language; and (3) the Chinese region of Vietnam.

There are no examples of Chinese literature written in Vietnam while it was under Chinese rule (111 BC–AD 939); there are only scattered examples of Sanskrit inscriptions written in Cambodia and Indonesia; yet most of the

literary works produced at the court of Pagan in Burma (flourished *c.* 1049–1300) have survived because the texts were copied and recopied by monks and students. But in the 14th–15th centuries, vernacular literatures suddenly emerged in Burma and Java, and a "national" literature appeared in Vietnam. The reasons behind the development of each were the same: a feeling of nationalistic pride at the final defeat of Kublai Khan's invasions, the desire of the people to find solace in literature amidst change and struggles for power, and the lack of wealth and patronage to channel artistic expression into building temples and tombs. In Vietnam and Java literary activity centred on the courts; but in Burma the first writers were the monks and, later, the laymen educated in their monasteries. In the new Burmese kingdom of Ava (flourished after 1364), the Shan kings were proud of their Burmese Buddhist culture, and they appointed the new writers into royal service, with the result that courtiers became writers also. The Tai kings of Laos and Siam led their courts in learning Pāli from the Mons, whom they had conquered, and Sanskrit from the Khmers, whom they harassed; nevertheless, seized with national pride and influenced by the Burmese example, they developed their own vernacular literature. But Cambodia itself declined. Although the monks in the Theravāda Buddhist (*i.e.*, the Southeast Asian form of Buddhism) monasteries produced a few works in Pāli, no vernacular literature emerged until finally Khmer-speaking people (those living in the area comprised approximately of modern Kampuchea) were borrowing many words from the Tais.

For its vernacular literatures, Southeast Asia can be divided into (1) Burma; (2) Thailand, Laos, and Cambodia; (3) Vietnam; (4) Malaysia and Indonesia; and (5) the Philippines (which produced a vernacular literature only in the 20th century, after the imposed Spanish and English languages and literatures had made their impact).

The era of royal patronage

Prestige of the writer. During the time of the kings, a Southeast Asian writer enjoyed patronage and a prestigious position in society. He could not, however, make a living by writing as a profession. Manuscripts had to be written by hand, and only in the case of famous works might one or two duplicates be made, again by hand. There was no question of selling the manuscript. A writer could only hope to attract the notice of his king and obtain a monetary reward or a royal office. By the time that printing presses were introduced, in the colonial period in the 19th century, the kings were gone and with them their writers. Colonial rule overwhelmed and destroyed vernacular literary traditions, leaving intact only oral literatures in the forms of folktales and folk songs. Literary criticism, as understood in Western cultures, had never been known, either in the ancient or modern literatures of Southeast Asia. Apart from a few stray writings on versification, therefore, no works of literary criticism or literary history existed until the colonial period. Even then, the interest of European scholars was chiefly confined to archaeology, and only a few made the attempt to study some special type or period of a vernacular literature (for example, vernacular versions of the *Rāmāyaṇa*, the great Sanskrit epic of India, or of 14th-century Javanese verse). There is a work in French dealing with Thai literature and a work in Burmese dealing with Burmese literature; but apart from these no study of any Southeast Asian literature as a whole has yet been made. For this neglect, native scholars are much to blame. Works on literary criticism and the history of literature would help give perspective to indigenous writers in the 20th century, both to those who want to cling to their native traditions (as is the case with writers in Burma and Thailand) and those who want to make a complete break with the past (as is generally the case in Vietnam, Indonesia, and Malaysia).

PRE-EUROPEAN COLONIAL PERIOD

Burma. The Burmese borrowed many words from Pāli but not to the extent that the Indonesians, Khmers, and the Thais borrowed Sanskrit words. The Burmese language was monosyllabic and tonal, and since there was no accent or stress, the feature that distinguished verse from prose was the regular occurrence of rhyme. They modelled their literature not on classic examples from Pāli or Sanskrit but on their own traditional folk songs.

The 15th century. In the 15th century, four types of verse existed: (1) *pyo* (religious verse), which retold stories of Buddha's birth and teaching and were taken from the *Jātaka*s (a collection of folktales adapted to Buddhist purposes and incorporated into the Pāli canon), to which were added imaginative details and a Burmese background; (2) *linkar* (shorter religious verse), or a devotional poem, characterized by a metaphysical flavour comparable in many ways to that which informs the work of the early 17th-century English poets George Herbert and Robert Herrick; (3) *mawgoon* (historical verse), half ode, half epic, written in praise of a king or prince and developing out of military marching songs; (4) *ayegyin* (lullaby), an informative poem usually addressed to a young prince or princess and written in praise of his royal ancestors.

Literature in the 15th century is dominated by three monks: Shin Maha Rahta Thara, who wrote for the court of Ava, and Shin Maha Thila Wuntha and Shin Uttamagyaw, both of whom were of village stock and did not go to court but remained on in their village monasteries. Shin Maha Thila Wuntha, in the closing years of his life, turned to prose and wrote a chronicle history of Buddhism. In this period several courtiers, both men and women, also began to achieve some literary success, and the genre called *myittaza* (epistle) first evolved, which is a long prose letter written by a monk and addressed to the king to advise him of his duties.

Poets at court

The 16th century. In the 16th century, the Burmese conquered Siam, and their subsequent knowledge of Thai romantic poems gave rise to a new verse form called the *yadu* (the seasons). They borrowed only the theme, however, and not the form, and they developed it as an emotional poem, passionate, yet with something of the cool intellectual strength of the poems of the English metaphysical poets John Donne and Andrew Marvell. The most famous writers of the *yadu* were two court poets, Phyu and Nyo; a general of the army called Nawaday; and Natshinnaung, king of Toungoo. The wide popularity of the poems eventually gave rise to a mock-heroic form called *yagan* ("Kick the *yadu*").

Golden age of literature. In the early years of the 18th century, U Kala compiled a history of Burma, written in precise and clear prose; the closing years, which coincided with the establishment of the third Burmese Empire, saw a great period of literature. The Thai court, brought as captives to the Burmese capital, introduced to the Burmese poetic romances and their *Rāma* play (based on the *Rāmāyaṇa*). Contact with the Thais stimulated the growth of a Burmese court drama and led to the appearance of Burmese court romances in poetic prose. The king's treasurer, however, made fun of the Thai importations and wrote the *Rama Yagan*, in which the high romance and courtly elegance of the 4th-century-BC *Rāmāyaṇa* ("The Life of Rāma") were given a rustic setting, with hilarious results. From the quiet of their monasteries, the monk Awbatha wrote novel-like rendering of the *Ten Long Jātakas* and the monk Kyeegan Shingyi wrote homely, pithy, and sometimes even humorous myittaza ("epistles") from villagers to their relations in the cities.

The defeat suffered by the Burmese in the Anglo-Burmese War of 1824–26—their first defeat since the time of Kublai Khan in the 13th century—introduced a note of melancholy to Burmese literature. During this first half of the 19th century, many of the new melancholy lyrics were set to music. Two great writers were a product of this period: the dramatist U Kyin U and the courtier U Pon Nya, the greatest writer of the time, whose plays, epistles, and songs are full of humour and zest for life.

Thailand. Until 1824 Thai literature was entirely the province of the king and his court: the king maintained a corps of writers, and it was the custom to attribute authorship of any literary work to the king himself. Thai vernacular literature began with verse, based on Sanskrit models but relying on an elaborate rhyme scheme because the

Elaborate rhyme scheme of Thai poetry

Siamese language was tonal. The two earliest known poems were *Yoon Pai* ("The Defeat of the Yoons"), an epic-ode having similarities to the Burmese *mawgoon* genre, and *Mahajati* ("The Great *Jātaka*"), a poem stressing ethical ideas, similar in form to the Burmese *pyo*. Both poems, written during the period 1475–85, give ample proof that Thai writers, using Sanskrit, Khmer, and Burmese models, could nonetheless produce a truly Thai work.

First golden age: King Narai (1657–88). All literary activity ceased in the 16th century because of the unsettled conditions that prevailed before and after the annexation of the country by the Burmese. Independence was regained toward the close of the century, and under King Narai (1657–88), at his court in Ayutthaya, Siamese literature achieved its first golden age. Narai was himself a great poet, and during his reign new verse forms were evolved. He wrote poetic romances, based on stories from the "Fifty Jātakas," which were in fact folktales belonging to the region retold in Pāli and disguised as *Jātaka*s by an unknown Tai monk. Narai also wrote the final version of the poem of tragic romance, *Pra Lo* ("Lord Lo"), which had first been composed by an anonymous author in a much earlier reign. Among courtier poets of this time, the most famous were Maharajaguru; Si Prat, a wild young gallant who wrote the romantic poem *Aniruddha* (the name of the hero of the poem) and some passionate love songs; Khun Devakavi, author of cradle-songs using many Sanskrit and Khmer words but modelled on the Burmese *ayegyin;* and Si Mahosot, the author of an ode-epic in praise of King Narai. A new genre, the travel poem, also became popular; and the first versions of the plays *Rāma* and *Inao* (based on Hindu–Khmer–Javanese models) were composed by the King and his corps of writers. Perhaps the only prose work of the period was the *History of Ayutthaya* by Luang Prasroeth, which was lost and came to light only in the 20th century. It showed some signs of being influenced by U Kala's *History* (of Burma).

Second golden age: King Rama II (1809–24). Siam was conquered by the Burmese in 1767, and a new dynasty was established in a new capital, Bangkok. Some effort was made to revive the country's culture, largely destroyed following the sack of the old capital of Ayutthaya; and under the poet-king Rama II a second golden age of Thai literature occurred, during which women achieved prominence as poets for the first time. The King, with his writers, composed the final versions of *Rāma* and *Inao* and also a popular romance, "Khun Chang and Khun Pen," based on an incident in Thai history. The most famous poets were Prince Paramanuchit, whose ode-epic *Taleng Phai* ("The Defeat of the Mons") testified to his greatness, and Sunthon Phu, the King's private secretary, who was born of humble parents but made his way in the court by the excellence of his poetry. A strongly religious king, Rama III disbanded the corps of writers and discouraged the performance of plays at his court. Sunthon Phu lost his position but wrote his most famous poem, *Phra Aphaimani,* away from the court. A long fantasy-romance, this work can be regarded as the end of court domination in literature. Further, a royal official composed a Thai translation in prose (*Sam Kok*) of the Chinese classic *Romance of the Three Kingdoms.* The author, Pra Klang, was admittedly a royal official; nevertheless, the work was meant for the people rather than the court. It was followed by a spate of imitations and finally resulted in the development of the historical novel.

Final version of *Rāma*

Laos, Cambodia, and Vietnam. Laotian literature was in many respects a dialect branch of Tai literature, and, as in Thailand, it was the creation of the royal court. A number of popular romantic poems and prose lives of famous monks were composed, but their authors were unknown; all works, in fact, were by custom written anonymously.

The kings of Cambodia, fallen from high estate and often mere vassals of Thailand, could not inspire the rise of a vernacular literature. Only in the monasteries was there any literary activity, and this was written in the Pāli language.

In Vietnam, the emperors of the Tran dynasty (13th–14th century) were themselves poets and patronized a new literature—which, nevertheless, was still written in Chinese and was therefore national rather than vernacular. The writings themselves, however, were by no means a mere branch of Chinese literature. The country was afterward conquered once more by China and it was not until it regained independence that, under the patronage of the Le dynasty emperors (15th–16th century), a new age of literature began. Although the Chinese language was still used, some writers were beginning to use the vernacular (employing Chu-nom script, consisting of modified Chinese characters). Nguyen Trai, Emperor Le Thanh Tong, and Nguyen Binh Khiem were the great poets of this period. In 1651 Father Alexandre de Rhodes, a Roman Catholic missionary priest, invented a new romanized script (Quoc-ngu) that became the national script. Literature then began to reach the common people.

The national script

Literary works written before the end of the 18th century have not survived; the best known are those written in the 19th century, before the country became a French colony in 1862. Ho Xuan Huong, Nguyen Cong Tru, Chu Manh Trinh, and Tran Ke Xuong were famous court poets. Nguyen Du (1765–1820) wrote moral tales in verse that appealed not only to the court but to the common people. His most famous work was *Kim Van Kieu,* a poem of 3,253 lines, showing a strong Chinese influence (the plot was taken from a Chinese historical novel, and its ethical basis was both Confucian and Chinese Buddhist). The plays of the period, although written in Vietnamese, followed Chinese dramatic traditions because the Vietnamese theatre was still Chinese in style and practice.

Malaysia and Indonesia. Malaysia and Indonesia together have about 300 different languages and dialects, but they have a single common linguistic ancestor. Before the coming of Islām to the region in the 14th century, Javanese had been the language of culture; afterward, during the Islāmic period, Malay became the most important language—and still more so under later Dutch colonial rule so that, logically, it was recognized in 1949 as the official Indonesian language by the newly independent Republic of Indonesia.

During the period of Indian cultural influence, Sanskrit flourished in the great empires that included both the Malay Peninsula and the islands of present-day Indonesia. In the 11th century, at the court of Emperor Airlangga, a national literature (as distinct from a vernacular literature) emerged. It was written in courtly Javanese mixed with Sanskrit words, and it used Sanskrit metres and poetic style. In the 14th century in Majapahit (the new Javanese Empire that had been established after the final defeat of Kublai Khan's forces) a vernacular literature based on the speech of the common people came into being. The most important work of this new literature was *Nāgarakertāgama* (1365), a long poem in praise of the king (though it was not a product of the court) that also contained descriptions of the life of the Javanese people at the time. Although it employed a number of Sanskrit words, the style and metre were Javanese, not Sanskrit.

The Indian Hindu epics had already been popularized in the Malay Peninsula and in the islands of Indonesia (by way of the shadow-puppet play), and in this period fresh versions began to be written in the new Javanese. Romances, called *hikayat,* both in verse and in prose, also appeared—having as their source native myth and legend. Soon Malay, Balinese, Sundanese, and Madurese vernacular literatures emerged, all dealing with the same themes.

Romance themes from myth and legend

The coming of Islām coincided with the rise of Malacca and the decay of Majapahit; but the popular fantasy-romances were able to survive by adopting a Muslim, instead of a Hindu, guise. New romances, telling the stories of heroes known to Islām, such as Alexander the Great, Amīr Ḥamzah, and Muḥammad ibn al-Ḥanafiah, were added to their number, and translations of Persian Muslim stories and of works on Muslim law, ethics, and mysticism further enriched Malay literature.

The finest work of all in the Malay language was the *Malay Annals,* written in about the 15th century. It gave a romanticized account of the history of the kingdom of Malacca and a vivid picture of life in the kingdom. Although a court record that begins with ancestral myths, it goes on to describe latter-day events of the kingdom with realism and humour.

In the Malay Peninsula, the coming of colonial rule did not at once overwhelm the existing native literature. As at the courts of the sultans of the British federated Malay states, the old traditions continued for some time. In Indonesia, however, a complete break was made with the cultural tradition.

EUROPEAN COLONIAL AND MODERN PERIODS

Impact of colonial rule

The entire region of Southeast Asia, with the single exception of Thailand, fell under colonial rule, and Thailand itself survived more as a buffer state than as a truly independent kingdom. At the courts of the kings of Laos, Cambodia, and Vietnam, which fell under French suzerainty, and in the palaces of the sultans of the British Malay states, vernacular literatures managed to survive for a time; but since these literatures had long ago ceased to develop—as a result of harassment by the Thais in the case of Laos and Cambodia, by the Portuguese in the case of Malaya, and by the French in the case of Vietnam—they soon became moribund. In all of Southeast Asia, except Burma and Thailand, the vernacular languages themselves lost their status, as the languages of the colonial rulers became the languages of administration and of a new elite. A revival of interest in the native languages and literatures occurred only toward the close of the colonial period, as a consequence of national movements for freedom.

Burma. In Burma, unlike India and other parts of the British Empire, English did not fully replace Burmese as the language of administration. In the almost classless Burmese society the language of the court and of literature was also the language of the people, which prompted the British government to retain Burmese as a second official language and to make both languages compulsory for study in schools and colleges. As a result, no English-speaking elite emerged, English literature did not dazzle native scholars, and, although its growth was retarded, Burmese literature did not disappear. With the intensification of the movement for freedom, around 1920, political tracts, novels, short stories, and poems reflected a political bias against colonial rule. In 1930, at the University of Rangoon, a group of young writers developed a new style of Burmese prose and poetry, a style little influenced by Western literature. In the post-independence period, novels and poems became centred on biographical and historical writings.

Thailand. Administrative and educational reforms introduced by King Mongkut (1851–68) as an answer to the threat of colonial conquest created a liberal atmosphere and a new reading public, and soon many of the old courtly writings were popularized in the form of romantic prose fiction. About 1914, King Vajiravudh, a graduate of Cambridge University, attempted to win back for the palace the leadership in literature; although he produced some fine adaptations of Shakespeare's plays, they made no impact on the people, with whom romantic fiction remained popular. Because of increased contact with the West, after World War II novels and short stories based on western models began to rival the earlier prose romances.

Laos, Cambodia, and Vietnam. Because of France's restrictive colonial educational policy, French language and literature never reached the common people. Moreover, the French-speaking elite, engrossed in French literature, neglected the native literature. With the growing vehemence of the freedom movement in the 1930s, however, there developed in Vietnam a new school of vernacular poetry that was less traditional and more nationalistic. But in the turbulent years that followed, the poets, including Ho Chi Minh himself, became occupied more with war than with literature.

Malaysia and Indonesia. The first Malaysian newspaper in the vernacular language, which appeared in 1876, introduced a new style of prose, less literary and nearer to the spoken Malay. Becoming immensely popular, the new style was further developed by other newspapers. (Although the early innovators were influenced by the English language, their followers were influenced by Arabic.) In about 1920 this new "Malaysian Malay" finally replaced the old literary Malay. The Translation Bureau, established by the British government in 1926, translated a great number of English books into the new Malay. In Indonesia, also, the old cultural language, literary Javanese, ceased to be used; by the end of the 19th century young Indonesians, overwhelmed by Dutch literature, started to write in Dutch. For example, a young girl, Raden Adjeng Kartini, wrote in Dutch a remarkable series of letters, containing criticism of Indonesian society, that were later collected and published; and a group of young men wrote poems in Dutch, although with an Indonesian background. By about 1920, however, the Dutch government itself had decided for political reasons to discourage further development of a national literature in Dutch, and the nationalist leaders had become eager for a new literature in the native language. This common aim bore fruit in 1933, when a literary journal under the editorship of Takdir Alisjahbana appeared, containing poems and essays written by various authors in the new Malay, which they now called Indonesian. The editor himself later wrote in Indonesian a number of popular novels containing social criticism, which were imitated by other writers. During the Japanese occupation of Indonesia and Malaya, this new Indonesian literature became popular also in Malaya. The adoption of Bahasa Malay (Indonesian) as the official language of Indonesia in 1949 gave further impetus to the development of the vernacular literature in both countries. The new tradition developed after independence, and its outstanding writers in Indonesia were, in poetry, Chairil Anwar and Sitor Situmorang. Important novelists include Ananta Toer and Takdir Alisjahbana.

Malaysian Malay

The Philippines. Philippine literature had its beginnings in great epics that were handed down orally from generation to generation and sung on festive occasions. When the Philippines became part of the Spanish Empire in the 16th century, printing was introduced, and all the early published works in the vernacular (Tagalog) were of Christian religious subjects. Eventually, some individual romantic legends taken from the epics were published, but they had acquired a European flavour. An outstanding work in the early years of the 19th century was an epic romance called *Florante at Laura* by the first native writer to achieve prominence—Francisco Balagtas—who wrote in Tagalog. In the latter half of the 19th century, an intellectual renaissance coincided with the beginnings of a national movement toward freedom; writers began using Spanish, for their work was part of the nationalist propaganda. The most famous author was José Rizal, who wrote a series of brilliant social novels, beginning with *Noli me tangere* ("Touch Me Not"). Other prominent writers, all essayists, were Mariano Ponce and Rafael Palma. There were poets also—for example, José Palma, whose poem "Filipinas" was later adopted as the national anthem. After the United States had taken over the Philippines, Spanish was gradually replaced by English, and new writers began to use that language as their medium. But before a new national literature could evolve, World War II took a heavy toll of writers, and those who survived became caught up in the political changes that followed. Many still write in English—the Spanish tradition, too, remains strong—but more and more writers are turning to Tagalog for literary expression. (M.H.Au.)

19th-century intellectual renaissance

Music

GENERAL CHARACTERISTICS

Society and music. *Rural and urban music.* A general musical division exists between the urban and rural areas of Southeast Asia. Urban centres comprise the islands of Java and Bali and places in Thailand, Laos, Kampuchea, and Burma, where big ensembles of gong families play for court and state ceremonies. Rural areas include other islands and remote places, where smaller ensembles and solo instruments play a simpler music for village feasts, curing ceremonies, and daily activities. In cities and towns influenced by Hindu epics such as the *Rāmāyaṇa* and *Mahābhārata,* shadow and masked plays and dances utilizing music play important communal roles, while in less urbanized areas, in lieu of musical plays, chants and songs in spirit worship and rituals are sung in exclusive surroundings—a ritual procession on the headwaters of

Borneo, a drinking ceremony in the jungles of Palawan, a feast in the uplands of Luzon.

In both regions the physical setting is usually the open air—in temple yards and courtyards, under the shade of big trees, in house and public yards, fields and clearings. Many musical instruments are made of natural products of a tropical environment, and their sounds are products of this milieu. The music of buzzers, zithers, and harps is thus akin to sounds heard in the tropical vegetation of Southeast Asia. In Bali, for example, special ways of chanting and sounds of the jew's harp (also jaw's harp) ensemble (*genggong*) imitate the croaking of frogs and noise of animals.

Relation to social institutions. Music in Southeast Asia is frequently related to ceremonies connected with religion, the state, community festivals, and family affairs. In Java, important Islāmic feasts, such as the birthday of Muḥammad or the fast of Ramaḍān, as well as animistic ceremonies marking the harvest and cycles of human life, are celebrated with shadow plays (*wayang* [*wajang*]). In Bali, the *gamelan gong* orchestra opens ceremonies and provides most of the music for temple feasts. The *gamelan selunding,* an ensemble with iron-keyed metallophones (like xylophones but with metal keys), plays ritual music, and the *gamelan angklung,* so called because it formerly included tube rattles, or *angklung,* is used to accompany long processions to symbolic baths near the river.

In Malaya the court orchestra, or *nobat,* was held almost as sacred as the powers of the sultan himself. Among the Land Dayak and Iban in Borneo, ceremonial chants are sung in feasts related to rice planting, harvesting, and honouring the omen bird *kenyalang* and other spirits.

The relation of music to dance and theatre. In the Thai masked play, or *khon,* dancers, chorus, soloists, and orchestra are all coordinated. The musicians know the movements of classical dance and coordinate musical phrases with dance patterns, turns, and movements. In the shadow play, or *nang sbek,* the dancer, who manipulates a leather puppet, must keep his foot movements in time with vocal recitations. During pauses in which the gong ensemble plays an interlude, the dancer must change steps accordingly. In general, when there is solo singing, the instrumental ensemble remains silent or plays only a few instruments in contrast to interludes of acrobatic shows or scenes of fighting, when the full orchestra clangs on all the instruments. In Balinese dancing, body movements, paces, and directions are dependent on drum strokes and signals from a wood block (*keprak*) and cymbals (*tjengtjeng*). The dancers generally rehearse with the musicians to know exactly when choreographic changes take place.

Musicians' knowledge of the dance

As theatre, the stories of *Rāmāyaṇa* and *Mahābhārata* have different musical supports, depending on the country. In Bali, *Mahābhārata* shadow plays are presented to the accompaniment of a quartet of metallophones known as *gender wayang.* In Cambodia, where the preference is for stories of the *Rāmāyaṇa* (which is called *Ramker* in Cambodia), the music is a full gong ensemble similar to the Thai *pi phat* ensemble, while in Burma, a percussion orchestra of drums and gongs in circular frames accompanies singing, dancing, and dialogues in all types of plays.

Musical traditions and practice. *Vocal music.* The role of the voice in music making differs from that of European music in both concept and execution. Men's and women's voices are each not divided into high and low ranges but are used for their colour qualities. In the Javanese shadow play, for example, the narrator (*dalang*) assumes many singing and speaking qualities to depict different characters and scenes. Arjuna, the chief *wayang* hero, is represented with a clear voice, speaking in a single tone. Puppets with bigger bodies are given lower, resonant voices. In Thai masked plays there is no desire to produce full open tones, as in Italian bel canto. A vocal tension accounts for shades of "nasal" singing that can be discerned in commercial recordings of Thai, Javanese, Cambodian, and Vietnamese music. In the Javanese orchestra (gamelan) the voice tries to imitate the nasality of the two-stringed fiddle (*rebab*). In Bali, a particular use of men's voices is in the *ketjak,* a ritual in which groups seated in concentric circles combine markedly pronounced syllables into pulsing rhythmic phrases. In village settings among the Kalinga of Luzon, in the Philippines, singing, speaking, or whispering of vowels is so subtle as to blur the border line between speech and song. On the Indonesian island of Flores, leader–chorus singing, with the chorus divided into two or more parts, is accompanied by a prolonged note (drone) or by a repeated melodic, rhythmic fragment (ostinato). In Borneo, or Mindanao and Luzon in the Philippines, a man or woman may sing an epic or a love song in a natural voice with little or no attempt to nasalize it. Epic singing, with long or short melodic lines, goes on for several nights, and some of the sounds are mumbled to give words and their meanings a particular shading. Further, a sensuousness in the quality of Islāmic singing is achieved through the use of shades of vowel sounds, vocal openings, and a bell-like clarity of tones.

Voice quality and dramatic meaning

Instrumental music. Although gong orchestras consisting of gongs, metallophones, and xylophones bind Southeast Asia into one musical cultural group, the types of ensembles and sounds they form may be classified into four areas. Java and Bali make up one unit because of their predominant use of bronze instruments in orchestras that make one homogeneous sound. Thailand, Laos, and Cambodia form another subdivision, with families of musical instruments producing heterogeneous sounds: the bronze group makes slowly decaying sounds, wooden xylophones play short sounds, and a reed blows a penetrating melody accompanied by a fourth group of cymbals, drums, and another gong. The Burmese orchestras differ from the Indonesian and Thai groups by the unique use of a row of tuned drums (sometimes called a drum circle), with sounds consisting of sharp attacks and quick-vanishing waves. The fourth area, Indonesia, Malaysia, and the Philippines, uses several types of suspended and horizontally laid gongs. These gongs produce various combinations of sounds. In Nias, an island west of Sumatra, one group of three heavy suspended gongs plays three rhythms of homogeneous sounds. Suspended gongs with a wide rim and a high knob are played alone, with another gong or with a drum on the Philippine islands of Mindanao and Palawan and the Indonesian island of Kalimantan (Borneo). Gongs laid in a row, called *kulintang,* are melody instruments accompanied by a percussion group. The most developed melodies are found in Mindanao, and the area of distribution extends to Borneo, Sumatra, and Celebes, in Indonesia. The sets of tuned gongs found throughout Southeast Asia are also called gong chimes, gong kettles, and gongs in a row.

Homogeneous and heterogeneous sound

Tonal systems. In contrast to the Western diatonic-scale system (based on seven-note scales comprised of whole and half steps) and its association with relatively "fixed" pitches, there prevails a gapped system in Southeast Asia (*i.e.,* scales containing intervals larger than a whole step) with elastic intonation. Examples include the five-tone *slendro* and the seven-tone *pelog* of Java and the seven-tone scale of Thailand. In each of these systems the distances between corresponding tones in two different sets of octaves are not exactly the same. For example, one Javanese *slendro* octave has the following intervals expressed in cents (a unit of pitch measurement; 1,200 cents make 12 semitones or one octave): 246, 241, 219, 254, 246; another has 245, 237, 234, 245, 267. In contrast, two tunings of the Western chromatic scale theoretically always have 12 semitones of 100 cents apiece.

Modal systems

Related to tonal systems are modes, which in Southeast Asia use tones of a particular scale system to form melodies. Associated with a given mode are a hierarchy of pitches, the principal and auxiliary tones, endings of melodic phrases (cadential formulas), ornaments, and the vocal line. Modes express emotions and are applied to different times of the day and night and to particular situations in stage plays. They are clearly present, with local variations, in Java, Vietnam, and Burma but are less distinct in Bali, Thailand, Laos, and Cambodia.

In rural areas a multitude of scales with mixed diatonic and gapped systems and no modes are used.

Musical time and improvisation. Musical time is generally divisible in units of two or four in urban music, but it occurs more freely and without a metric pulse in rural areas, especially in singing. Musical improvisation or the use

of variations based on a melodic theme is not universal. It is essential to the playing of the *rebab* and singing in the Javanese gamelan, the tappings on the Burmese circle of drums, and the percussive playing on the *kulintang*. But, in fast playing in the Balinese gamelan, exact repetitions of patterns are necessary, for there is no time for the performer to think of alternative formulas. Similarly, the separate rhythmic patterns of five instrumental parts do not change in the gong (*gangsa*) music of the Ibaloi of Luzon. Repetition is the essence of the music.

HISTORICAL DEVELOPMENTS

Origins. *Early bamboo instruments.* The widespread use of bamboo musical instruments in practically all parts of Southeast Asia points to the antiquity of these instruments and, probably, that of the music they play. A historical citation of mouth organs and jew's harps in the Chinese *Shih Ching* ("Classic of Poetry") shows that these instruments were known in the 8th century BC. Previous to this time, other bamboo musical instruments were probably in use, just as bamboo tools were used in pre-Neolithic times.

Drones and ostinato

The music of pre-Neolithic types of bamboo musical instruments, such as are played in the 20th century, may be just as old as these instruments. One general feature that points to this antiquity is the widespread and frequent use of a very simple musical element: a sustained tone (drone) or repetition of one or several tones (ostinato). Sustained tones appear in the mouth organ, where one or two continuous sounds are held by one or two pipes while a melody is formed by the other pipes. Prolonged tones may also be heard in rows of flutes played by one person in Flores. One flute acts as ostinato and the rest make a melody. In group singing, an underlying held tone is common. Repetition of tones occurs in bamboo instruments (jew's harps, percussion tubes and half percussion tubes, zithers, clappers, slit drums) as well as in nonbamboo instruments. In the *kudjapi*, a two-stringed lute, one string is used for the ostinato and the other to pluck the melody. In the log drum, two players play fast rhythms of continuous sounds while another player taps improvised rhythms.

Bronze instruments in gong families of Indonesia, Thailand, and Burma employ repeated sounds acting as ostinati. A widespread and preponderant use of dronelike or repeated sounds in Southeast Asia shows that they are probably an ancient fundamental musical element.

Early bronze instruments. The earliest bronze musical instruments are kettle gongs (deep-rimmed gongs), which date back to about 300 BC and are found in Vietnam, Bali, Sumatra, Borneo, Thailand, and Burma. In Burmese gongs the use of a heavy beater for the centre and a lighter stick to strike the side denotes an opposition of a full and a tiny sound applied today also to the *babandil* and other gong ensembles in Palawan and Borneo.

Gongs that predominate in Southeast Asia are those with a boss, or central beating knob. The many varieties differ according to their shapes, chemical properties, playing position, number in a series, manner of playing, musical function, and sound. Flat gongs without a central boss are not as widely used. They are found in the hills of Thailand, Laos, Cambodia, Vietnam, in some parts of Indonesia, and in the northern Philippines and may have come to Southeast Asia either through China in the 6th century or from the Middle East.

Musical traditions. The influence of the great traditions of Asia—Indian, Chinese, Islāmic, and Khmer (Cambodian)—on native Southeast Asian music varies in different countries. From India come principally two ancient Sanskrit epics—the *Mahābhārata* and the *Rāmāyaṇa*. Deep attachment to themes from the *Rāmāyaṇa* pervades the whole Southeast Asian region, except the Philippines, where Indian influence was weakest. Musical instruments attributed to India and appearing in 9th-century reliefs at the Buddhist temple of Borobuḍur and Hindu temple of Prambanan, in Java, are bronze bells, bar zithers, cymbals, conical drums, flutes, shawms, and lutes. They may still be found in several islands of Indonesia. Khmer gong circles, stringed instruments, mouth organs, drums, and oboes still in use in rural Cambodia and Vietnam are depicted in the 12th-century ruins at Angkor Wat in Cambodia. Prehistoric lithophones, or stone chimes, excavated in Vietnam in 1949, may have been the ancestors of kettle gongs. Chinese-type musical instruments (two- and three-stringed fiddles, bells, and drums), the use of the Chinese pentatonic (five-tone) scale, and duple and quadruple time (typical Chinese metres) are used in Vietnam, Burma, Thailand, Laos, and Cambodia. Islāmic musical instruments—drums, two-stringed fiddles (*rebab*), and three-stringed lutes—may be heard in Java, while melismatic singing (many notes to one syllable), especially in Islāmic rituals, is usual among the Malay groups on Borneo.

Mingling of influences

There are also musical instruments and elements that have developed locally. The mouth organs of Borneo, Laos, and Cambodia are probable ancestors of the Chinese *sheng* and the Japanese *sho* (mouth organs). Jew's harps, tube zithers, ring flutes, buzzers, xylophones, two-stringed lutes, and various types of gongs with boss (knobbed centre) are some of the most typical instruments of Southeast Asia. A probably ancient manner of measuring flute stops in Mindanao—dividing flute segments into proportional lengths to produce the octave, fifth, and other intervals—recalls a very old Chinese account of cutting bamboo tubes into lengths that would sound these same intervals.

In general, music in Southeast Asia is a tradition taught to each succeeding generation without the use of written notation. From exclusive families of musicians in courts, gamelan music was transmitted to the people. Epic and ritual songs are learned by rote and handed down from older to younger generations. Hence, skill in instrumental music is developed by imitation and practice.

Burma. Just as today all types of Burmese plays are accompanied by the traditional Burmese orchestra, the beginnings of Burmese theatre contained a music that, like the theatre, was probably based on primitive religious rituals. Before Indian and Chinese musical influences, the inspirational source of Burmese music and dance was the miracle plays (*nibhatkhin*), which, in turn, were based on singing, dancing, and entertainment in local folk feasts that date back to antiquity. The worship of spirits (*nats*) at Chinese festivals was accompanied by women who, through song and dance, communicated with and were possessed by these spirits. Following this practice, professional entertainers taking the place of women danced, sang, and played instruments during the first *nibhatkhin*. These practices led to the dancing and singing associated with the *pwe*, a popular play for public and courtly entertainment.

Foreign musical influences came from India, China, and Thailand. Indian elements appear in musical terms, theories about scales, and in some musical instruments—oboe, double-headed drums, cymbals, and the arched harp. Chinese influence appears to be older and is apparent in the use of the pentatonic scale and such musical instruments as table zithers (related to the Chinese *ch'in*), a dragon-head lute resembling a Chinese *p'i-p'a*, and two- and three-stringed fiddles. From Thailand and the Khmer civilization of Cambodia probably came both the use of gongs in a circular frame and the dramatization of episodes from the *Rāmāyaṇa*. In the traditional orchestra for state ceremonies, for the theatre, and, formerly, for royalty, three simultaneous variations of the same theme are performed by two sets of melodic percussion—a circle of about 21 tuned drums (*saing-waing*) and a circle of about 21 tuned gongs (*kyi waing*)—and at least one oboe (*hne*) or a flute (*pulwe*). To this is added a playing of a percussion group comprising a double-headed drum (*patma*), a pair of cymbals (*la gwin*), and clappers playing a duple or a quadruple metre. In three rhythmic patterns applied by these percussion groups to specific song types, the strong beats are always marked by the clappers.

Melodic traits

Melodies played on traditional instruments (*saing-waing*, harp, *pattala* or xylophone) are frequently broken by rests and consist of segments of two, three, or four notes that form phrases, usually of eight or 16 beats. Several phrases make up a number of verses to complete a musical rendition. Melodies, based on modes, are constructed according to the previously discussed elements usually found in

the modal music of Southeast Asia. Song types exist in Burmese music and are assigned to specific modes.

The Burmese arched harp (*saung gauk*) has features that may be traced back to pre-Hittite times and the Egyptian 4th dynasty (*c.* 2613–*c.* 2494 BC). Scarcely existent outside of Burma, this instrument has undergone a renascence in the 20th century. A more popular solo instrument is a wooden xylophone *pattala.*

The following instruments may be found among ethnic groups in rural Burma: idiophones, or resonant solids—bamboo jew's harps, clappers, cymbals, wooden slit drums, bronze kettle gongs, drums; membranophones, or vibrating-membrane instruments—goblet drums; chordophones, or stringed instruments—crocodile zithers, monochords with calabash resonators, three- and four-stringed fiddles; aerophones, or wind instruments—lip-valley flutes, ring flutes, panpipes, double-reed winds, buffalo horns, and mouth organs.

Thailand, Laos, and Cambodia. Although their individual political histories differ, the music practiced in Thailand, Laos, and Cambodia is almost identical. The musical instruments and forms of this region spring from the same sources: India, the indigenous Mon-Khmer civilizations, China, and Indonesia. In Thailand, three types of orchestras, called *pi phat, kruang sai,* and *mahori,* exist. The *pi phat,* which plays for court ceremonies and theatrical presentations, uses melodic percussion (gongs in a circle, xylophones, metallophones) and a blown reed. The *kruang sai* performs in popular village affairs and combines strings (monochords, lutes, and fiddles with two and three strings) and wind instruments (oboes and flutes); while the *mahori,* as accompaniment of solo and choral singing, mixes strings (floor zithers, three-stringed fiddles, and lutes) and melodic percussion (gongs and xylophones) with the winds (flutes and oboes). All three ensembles are provided with a rhythmic group of drums, cymbals, and a gong to punctuate the melody parts. Some of the above musical instruments and their functions may best be illustrated in the *pi phat* ensemble below.

Thai ensembles

A slow-moving theme is played by gongs arranged in a circle (*khong wong yai*) with variations in smaller gongs (*khong wong lek*), two wooden xylophones (*ranat ek, ranat thum*), and two box-shaped metallophones (*ranat thong ek, ranat thong thum*). The last three pairs of instruments vary the theme by playing twice as fast or by repeating, anticipating, and revolving around it. A double-reed oboe (*pi nai*) hovers above the melodic percussion, providing the only blown sound in the ensemble. Together with the punctuating gongs and drums, the whole orchestra displays a polyphonic (many-voiced) stratification of instrumental parts, using unisons and octaves mainly in the strong beats.

A melody may be broken down into phrase units consisting of two or four measures that may be joined by four other phrase units to make a phrase block, and a given number of blocks constitutes one musical composition. Three speeds of rendition—slow, medium, fast—in either duple or quadruple time are marked by two alternating strokes in a pair of cymbals; a dampened clap marks a strong beat, and a ringing vibration denotes a weak beat.

The tuning system is made up of seven tempered (approximately equidistant) tones to an octave. But the melodies constructed out of this system use only five tones out of seven—which sound close to a Chinese pentatonic scale. This scale may be constructed in any of seven levels or tones of the Thai tuning system. Further, through a process called *metabole,* melodies may move from one level to another.

Shadow-play music and dance

In the Cambodian shadow play (*nang sbek*) two narrators alternate in chanted recitative to explain the role of the leather puppets. Dancers parading these figures across the screen and simulating their actions are accompanied by an orchestra. A limited number of tunes is played to eight dance positions (walk, flight or military march, combat, meditation, sorrow or pain, promenade, reunion, and metamorphosis). In the play these poses are assumed by princes, princesses, monkeys, demons, peasants, or ascetics.

Among different ethnic groups, such as the Khmer Saoch, Pwo Karen, Bu Nuer, Kae Lisu, Kuoy, and Samre, a rural music related to that of the ancient Khmer peoples is played by aerophones (buffalo horns, mouth organs, vertical flutes), idiophones (flat gongs, gongs with boss, cymbals, jew's harps), chordophones (bamboo zithers), and membranophones (circle of drums). Other important instruments for solo performance or as accompaniment to songs are the three-stringed crocodile zither (*chakhe*), a four-stringed lute (*grajappi*), a plucked monochord with a gourd resonator (*phin nam tao*), and a bamboo whistle flute (*khlui*).

Vietnam. Although Vietnamese music belongs to the great Chinese musical tradition, which includes the music of Korea, Mongolia, and Japan, some of its musical elements are indigenous or come from other parts of Southeast Asia, and some derive from Champa, an ancient Hinduized kingdom of Vietnam. Archaeological finds in the village of Dong Son revealed that the ancient Vietnamese used kettle gongs, mouth organs, wooden clappers, and the conch trumpet. From the 10th to the 15th century a joint Indian and Chinese element left its musical imprint. The Chinese seven-stringed zither (*ch'in*) and a double-headed drum were played together, or a Champa melody was accompanied by a drum. It was at this time that two traditional Chinese ensembles—Great Music and Little Music—and an elementary Chinese theatrical art were introduced. From the 15th to the 18th century the Chinese influence reached its height. Court music (*nha nhac*) was played by two orchestras. One, located in the Upper Hall of the court, consisted of a chime of 12 stones, a series of 12 bells, a zither of 25 strings (Chinese *se*), a zither with seven strings (Chinese *ch'in*), flutes, panpipes, a scraper in the shape of a tiger, a double-headed drum, a mouth organ, and a globular whistle. The second orchestra in the Lower Hall used 16 iron chimes, a harp with 20 strings, a lute with four strings (Chinese *p'i-p'a*), a double flute, a double-headed drum, and a mouth organ. Ceremonial music, almost nonexistent in the 20th century, was patterned after court music.

Traditional court music

In Buddhist ceremonies, prayers were recited in three ways: as recitation in a low voice, as a cantillation (sung, inflected recitation) following the six tones of the Vietnamese language, and as chant accompanied by an orchestra of two drums, bell, gong, cymbals, and fiddles.

Music as entertainment is mostly a vocal art played without ritual outside the court and still enjoyed by many people. The *hat a dao* found in the north is the oldest form. It is a woman's art song with different instrumental accompaniments, dances, a varied repertoire, and a long history of evolution.

From the 19th century to World War II, Vietnamese music reaffirmed its character. Although the playing of court music was restricted, popular music was encouraged, leading to northern and southern styles that were patronized by both the aristocracy and commoners. Western musical influence in this period was manifest in the use of the mandolin, the Spanish guitar, and the violin, as well as by the introduction of European classical music and composition following Occidental forms. In the late 20th century traditional Vietnamese music began to disappear, but attempts to revive it began in the early 1970s.

Vietnamese rural folk music is built on the same musical principles as court music. The main difference lies in its application to village activities—work, games, courting, marriage, cure for the sick, entertainment, feasts.

Elements of Vietnamese music

Common elements characterize and unify all Vietnamese music. It is based on an oral tradition, with written notation serving only as a reading guide. Melodies are generally built out of a pentatonic system (for example, C, D, F, G, A) to which two auxiliary tones (E, B) may be added to make other pentatonic melodies. A song, usually preceded by a prelude, may be sung in slow, moderate, or fast tempo divisible by two or four, with a simple contrapuntal (countermelody) accompaniment using unisons and octaves at beginning points of phrases. Outside of the first beats, intervals of fifths, fourths, thirds, and even seconds are allowed. An important aspect of melodies is the idea of mode (*dieu*), the elements of which do not essentially differ from those of Javanese and Burmese music.

Indonesia and Malaysia. *Java.* A Javanese philosophical concept based on mysticism, refinement (*halus*), and the inner life as related to Hindu, Islāmic, and Indonesian thought may best be represented in music by the Javanese gamelan, an orchestra made up mostly of bronze instruments producing homogeneous blended sounds. The instruments in the ensemble may be divided into three groups of musical function. The first group comprises thick bronze slabs (*saron demung, saron barung, saron panerus*) on trough resonators playing the theme usually in regular note values without ornamentation. The second group consists of elaborating or panerusan instruments, which add ornaments to the main theme. In this group gongs in double rows (*bonang panembang, bonang barung, bonang panerus*) play variations with the same ratio of speed as the *saron* group. In softer sounding music for indoor performance, other panerusan instruments with very mellow sounds come in. These are three sizes of thin bronze slabs with bamboo resonators—*gender panembung* or *slentem, gender barung,* and *gender panerus.* Other elaborating instruments are the wooden xylophone (*gambang*), the zither (*tjelempung*) with 26 strings tuned in pairs, an end-blown flute (*suling*), and a two-stringed lute (called a *rebab* by the Javanese), which leads the orchestra. In loud-sounding music, the soft-sounding instruments are not played, and the drum (*kendang*) leads the orchestra. The third group provides "colotomic," or punctuating beats in four rhythmic patterns played separately by four types of heavy, suspended, or horizontally laid gongs.

Instrument groups of the gamelan

Two tuning systems prevail. The *slendro* tends to have five equidistant but flexible (or varying) pitches in an octave, while the *pelog,* with seven equally flexible tones, has a more varied structure. One tuning with intervals expressed in cents (140, 143, 275, 127, 116, 204, 222) may roughly be represented by the following notes in a descending scale: C↑, A♯, G♯, G↓, F↑, D♯↓, C♯↑, and C. (Arrows up are tones slightly higher than Western tempered tuning [in which a semitone is equivalent to 100 cents] and vice versa for arrows down.) Melodies from these tunings are governed by a modal structure (*patet*) the elements of which are similar to those of Vietnamese and Burmese music.

Scales and modes

In West Java the most popular ensembles use a vocal part, a two-stringed fiddle (*rebab*) or a bamboo flute (*suling*), and a zither (*kachapi*). In the gamelan, submodes (*surupan*) are formed by the use of vocal tones—sung or played on the *suling* or *rebab*—which amplify the number of scales in both the *pelog* and *slendro* systems.

Bali. In contrast to the introspection of Javanese music, the Balinese gamelan exudes a music of brilliant sounds with syncopations (displaced accents) and sudden changes, as well as gradual increase and decrease in volume and speed and feats of fast, precise playing. The tuning system, musical instruments, and polyphonic stratification are similar to those of the Javanese gamelan, although in Bali the seven-tone *pelog* is not popular. Most gamelan are tuned to a five- or four-tone system, and the concept of modes is not as clearly developed as in Java. A variety of gamelan exists, each with a special function, instrumentation, repertoire, and tuning system. The *gamelan gong* orchestra is among the most extensive in its number of instruments. A modern version, *gong kebjar,* omits the *trompong* (gongs in a row) and *saron* (bronze slabs over a trough resonator) and replaces them with *gangsa gantung* (metallophone with bamboo resonators) and *rejong* of four gongs to produce exuberant outbursts of sound. The *gamelan gambuh,* now rare, comprises four end-blown flutes, one *rebab,* and a group of percussion. The *gamelan semar pegulingan,* played formerly in royal courts but now almost disappeared, emphasizes the *trompong* as a solo instrument. The *gamelan pelegongan* is a virtuoso orchestra that accompanies *legong* dances, while the *gamelan pedjogedan* is an orchestra of xylophones for dance (*djoged*) and entertainment in the marketplace. The *gender wayang* is a quartet of *slendro* tuned metallophones specially employed for shadow plays. The *gamelan angklung,* a village orchestra assembled during ceremonies, anniversaries, and cremations, originally consisted of rattling tubes that are now replaced by metallophones. The *gamelan ardja* is characterized by a soft timbre (tone colour) and the use of a one-stringed bamboo zither, the *guntang,* to accompany musical comedy and popular plays.

Varieties of Balinese gamelan

Other parts of Indonesia. In the islands of Flores, Nias, New Guinea, Celebes, and Borneo idiophones make up perhaps the most varied collection of musical instruments—gongs of various profiles, slit drums, jew's harps pulled with a string, clappers, bells, xylophones, percussion sticks, bull-roarers, and stamping tubes. Particularly interesting are idiophones made of bones, shells, skulls, fruits, seeds, planks, pellets, crab claws, clogs, coconut, and shark bones. Membranophones are represented by drums shaped like a cylinder, goblet, vase, round frame, hourglass, cone, cup, barrel, or a tube. Aerophones present an array of vertical and transverse (horizontally played) flutes, panpipes, ring flutes, shawms, clarinets, gourd trumpets, conch shells, ocarinas, and flutes with different mouthpieces. Chordophones include bamboo zithers, spike fiddles (in which the neck skewers the body), one- and two-stringed lutes, musical bows, monochords, guitars, *rebabs,* bar zithers, and sago zithers. In Flores, part singing with a sustained drone is frequent. Songs in Nias use diatonic (whole and half steps), chromatic (half steps), and gapped melodies largely less than an octave in range. In Sarawak descending melodies make up a tetrachord (four adjacent tones forming the interval of a fourth). In Indonesian New Guinea departures from songs with gapped scales include fanfare, stair descent, and tiled melodies (the last consisting of short phrases repeated at different pitch levels).

Malaysia. At least three principal cultural influences—Indonesian, Hindu, and Islāmic—left their musical marks in Malaysia. The Indonesian influence is seen principally in musical forms, participants, and paraphernalia of the Malaysian shadow play (*wayang kulit*). It is said that the Indian epics and, especially, the *Pandji* tales of Java came to Malaysia via Indonesia, but there are songs in certain plays and musical instruments (*e.g.,* the double-headed drum and oboe) that could have reached Malaysia from India through other routes. Islāmic traces are evident in melismatic songs among the Malay groups in songs connected with religious rituals and in choral singing in the *ma'yong* plays. Chinese music, a more recent development, is largely practiced among the Chinese communities, principally in Singapore.

Before Malaysian independence, the *nobat,* an old royal instrumental ensemble dating back to about the 16th century, played exclusively for important court ceremonies in the palaces of the sultans of Perak, Kedah, Selangor, and Trengganu. Today, in Kedah, the ensemble consists of five instruments: one big goblet drum (*negara*), two double-headed drums (*gendang*), one long oboe (*nafiri*), one small oboe (*nafiri*), and one gong. The music, which consists of ten surviving pieces, is broadcast today and performed live.

Three shadow plays exist, principally in the state of Kelantan. The *wayang gedek* is the Thai form; *wayang Djawa,* a Malay form, is almost extinct; and the *wayang Siam,* which is a combination of Thai and Malay influences, is the most popular form of puppet shadow play. The operator of the performance is the narrator (*dalang*), who manipulates the leather figures, introduces important characters, and describes different scenes with the accompaniment of the orchestra. The music is led by a two-stringed lute (*rehab*) in the *Rāmāyaṇa,* or an oboe (*serunai*) in *Mahābhārata* and *Pandji* cycles. The melodic instruments are supported by a percussion group consisting of pairs of goblet-shaped drums (*gedombak*), cylindrical drums (*gendang*), barrel drums (*geduk*), gongs lying on a support (*chanang*), suspended gongs (*gong*) or, sometimes, a row of gongs played by two or three men, and one pair of cymbals (*kesi*). The music usually begins with a prelude followed by a list of pieces the sequences of which are dictated by the narrator.

The *ma'yong,* a dance drama that probably dates back to more than 1,000 years, was introduced in Kelantan under the patronage of the royal courts. In the 20th century it exists as a folk theatre with an all-female cast. The music that accompanies 12 surviving stories is played by an orchestra of one bowed lute (*rebab*), two suspended

gongs, and a pair of double-headed drums (*gendang*). A heterophony (simultaneous variation of the same melody) between a solo voice, a chorus, and the *rebab* creates a music with a Middle Eastern flavour.

Variety in instruments

A rich musical heritage in the rural sections of Malaya is shown in musical instruments used by Malay, Thai, Semang, and Sakai groups. Idiophones include shell and coconut rattles, the jew's harp (mostly pulled by a string, rather than plucked), bull-roarers, bamboo clappers, and the bamboo slit drum. Aerophones include the buffalo horn, wooden and clay whistles, nose flutes, end-blown flutes, and the oboe. Chordophones are two- and three-stringed fiddles with coconut resonators, monochords, and tube zithers. One membranophone is a double-headed cylindrical drum.

In Borneo among the Malay, Kadazan, and Iban groups, the principal instruments are gongs in a row (*gulintangan*) played with suspended gongs of different types (*chanang, gong, tawak-tawak*). Among the Murut, Kenyah, and Iban the mouth organ with a calabash resonator (*sompoton*) plays a melody with a drone accompaniment. The jew's harp (*ruding*), bamboo zither (*tongkungon*), nose flute (*tuali*), hourglass drum (*ketubong*), and vertical flute (*suling*) may be heard among different ethnic groups. Iban ceremonial songs are sung in connection with rice festivals and rituals to prevent sickness, while mourning songs make up a rich repertoire of solo and leader–chorus singing. The Kenyah are particularly adept at blending low voices of men singing a melody supported by a drone.

The Philippines. Two musical cultures—Western and Southeast Asian—prevail in the Philippines. Western music is practiced by about 90 percent of the population, while Southeast Asian examples are heard only in mountain and inland regions, among about 10 percent of the people.

The Western tradition dates back to the 17th century, when the first Spanish friars taught plainchant and musical theory and introduced such European musical instruments as the flute, oboe, guitar, and harp. There subsequently arose a new music related to Christian practices but not connected with the liturgy. Processional songs, hymns in honour of the Blessed Virgin, Easter songs, and songs for May (Mary's month) are still sung in different sections of the country. A secular music tradition also developed. Guitars, string ensembles (*rondalla*), flute, drum, harps, and brass bands flourished in the provinces among the principal linguistic groups and still appear during town fiestas and important gatherings. Competing bands played overtures of Italian operas, marches, and light music. Young men, like their counterparts throughout the Hispanic world, sang love songs (*kundiman*) in nightly serenades beneath the windows of their ladies. It was not uncommon in family gatherings for someone to be asked to sing an aria, play the harp, or declaim a poem. Orchestral music accompanied operas and operettas (*zarzuelas*), while solo recitals and concerts were organized in clubs or music associations. With the advent of formal music instruction in schools, performance and composition rose to professional levels. In the 20th century several symphony orchestras, choral groups, ballet companies, and instrumental ensembles performed with varying regularity.

Gong ensembles

A Southeast Asian musical tradition exists completely apart from the Western tradition. In the north, flat gongs are played in different instrumental combinations (six gongs; two gongs, two drums and a pair of sticks; three gongs). In the ensemble with six gongs, four are treated as "melody" instruments, one as ostinato, and another as a freer layer of improvisation. The melody consists of scattered tones produced by strokes, slaps, and slides of the hands against the flat side of the gong. Other musical instruments in the northern Philippines are bamboo. These are the nose flute (*kalleleng*), lip-valley or notched flute (*paldong*), whistle flute (*olimong*), panpipes (*diwdiwas*), buzzer (*balingbing*), half-tube percussion (*palangug*), stamping tube (*tongatong*), tube zither (*kolitong*), and jew's harp (*giwong*). Leader–chorus singing among the Ibaloi is smooth and sung freely without a metric beat, while the same form among the Bontoc is emphatic, loud, and metric. Scales in songs and musical instruments use from two to several tones within and beyond an octave and are arranged as gapped, diatonic, and pentatonic varieties.

In the southern Philippines (the islands of Mindanao and Sulu), the more developed ensemble is the *kulintang*, which consists of eight gongs in a row as melody instruments accompanied by three other gong types (a wide-rimmed pair; two narrow-rimmed pairs; one with turned-in rim) and a cylindrical drum. The *kulintang* scale is made up of flexible tones with combinations of wide and narrow gaps sometimes approaching a Chinese pentatonic variety and oftentimes not. Its melody is built on nuclear tones consisting of two, three, or more tones to form a phrase. Several phrases may be built, repeated, and elongated to complete one rendition lasting two to three minutes. Pieces of music are played continuously for a long period during the night.

In the central west Philippines on the island of Mindoro, love songs are sung that are based on reciting tones with interludes played by a miniature copy of the Western guitar or a small violin with three strings played like a cello.

(Jé.Ma.)

The performing arts

In variety of dance and theatrical forms and in the number of performing groups, no area in the world except India and Pakistan compares to Southeast Asia. Some form of the performing arts is a normal part of life throughout the several nations. Sophisticated performing groups cluster in and around the present and former court cities—Jogjakarta and Surakarta in Java, Ubud and Gianjar in Bali, Bangkok in Thailand, Mandalay in Burma, Siem Reap near Angkor and Phnom Penh in Kampuchea (Cambodia), Hue in Vietnam—where drama, puppetry, dance, and music have been cultivated for ten centuries or more. Hundreds of commercial theatrical and dance groups perform in such newer centres as Rangoon, Saigon, and Jakarta and in scores of provincial cities and towns. Wandering troupes of actors, puppeteers, singers, and dancers travel from village to village in areas adjacent to these population centres. There are few communities in which some form of folk dance is not performed by local people.

Richness of form and experience

In the West, music, dance, and drama are usually separate arts, whereas in all areas of Southeast Asia, drama, dance, mime, music, song, and narrative are integrated into composite forms, often with masks or in the form of puppetry. The spectator's senses, emotions, and intellect are bombarded simultaneously with colour, movement, and sound. The result is a richness and a vividness in the theatre that is absent in most Western drama, so much of which rests on a literary basis.

More than 100 distinct forms or genres of performing arts can be distinguished in Southeast Asia. These can be grouped, according to which of the various stage arts is emphasized, into (1) masked dance and masked dance-mime, (2) unmasked dance and dance-drama, (3) drama with music and dance, (4) opera, (5) shadow-puppet plays, and (6) doll- or stick-puppet plays.

DIVERSE TRADITIONS IN THE PERFORMING ARTS

Four relatively distinct traditions exist in the performing arts: folk, court, popular, and Western.

The folk tradition. Dances in the folk tradition are exceptionally numerous and widespread. Some are performed as religious ritual, others, particularly on the Indonesian island of Bali, by highly trained and respected artists, and still another kind as entertainment in which the community participates. Folk theatre is more complex than folk dance and thus less widespread, but it has deep connections with religious ritual. Although the origins of most folk performing arts lie in remote times, later court forms exerted important influence on many of the folk forms. Conversely, folk forms have been a source of inspiration to court artists.

The court tradition. The shadow play and masked and unmasked dance are court arts reflecting centuries of subtle refinement under the patronage of kings and princes. In Southeast Asia the shadow theatre is a major classic art. Leather puppets of mythological figures, the bodies intri-

Shadow plays

cately incised to allow light to pass through, are attached to sticks for manipulation. A lacy shadow is created by a flaming lamp as the puppet is pressed against the back of a vertical screen of white cloth. The flickering and insubstantial shadow seen from the other side creates for the understanding viewer a mystic world with deep symbolic meaning. In Java, Bali, Malaysia, Cambodia, and Thailand shadow plays and their techniques have been emulated by human actors and dancers and have been the models for marionette and doll-puppet theatre.

Dance troupes have been a part of court life at least since recorded history began. In the mainland courts of Cambodia, Thailand, Laos, and Burma, concubines of the ruler's harem who performed female dances were segregated from male performers, giving rise to separate forms of female unmasked dance and male masked dance-mime. Although certain dances traditionally are performed only by men or only by women in Indonesia and Vietnam, mixed casts have a long history, especially in dramatic pieces. Court dance on the mainland and in Indonesia has been influenced by Indian dance style, and Vietnamese dance by the dance styles of Chinese opera, but they have acquired a distinctly Southeast Asian character. Court dance reached its greatest development when applied to mythological and legendary themes, often taken from the shadow theatre. The resulting dance-dramas and masked dance-mimes of Thailand, Cambodia, and Java are world famous for their magnificent scale and elegance of execution. Some of these court arts are no longer performed, and others face increasing difficulty securing financial support, yet they remain important.

The popular and Western traditions. In the popular traditions are those 400 to 500 professional troupes who perform, except in the Philippines, in commercial theatre buildings of major cities for an urban ticket-buying audience. Some forms of popular theatre are directly modelled on court dance-drama, but most are spoken drama in which court-derived music, song, and dance movements have been inserted. Local legend and history provide the subject matter for many of these plays. As in much of Asia, the performer in the popular tradition is seldom accorded status and may be despised as a vagabond.

The spoken drama, the ballet, and the modern dances are known only superficially in Southeast Asia. The sole exception is the Philippines, where amateur performances of Western plays constitute the country's main theatrical tradition. Southeast Asian audiences generally find Western plays based mainly on dialogue to be uninteresting and deficient in artistic qualities. European and American films and television programs, however, are widely shown and appreciated, and popular Western dances are found in major urban areas. Undoubtedly the impact of these forms on local audiences will continue to increase, possibly to the detriment of the indigenous traditions.

CHARACTERISTICS OF DANCE

Dramatic and nondramatic forms. In the parts of Southeast Asia influenced by Indian forms—everywhere except for Vietnam and the Philippines—nondramatic and dramatic dance are both known. Nondramatic, or "pure," dances that do not express emotional states of characters are numerous in both folk and court traditions. Among court dances, the Javanese *bedaja* is typical. Nine dancers move in unison, without emotional expression, in precisely fixed choreographic patterns designed to demonstrate sheer grace of movement. The *maebot,* composed as a Thai "alphabet of dance," is used to train pupils in the basic movements of court dance. Other dances that include character impersonation yet are not explicitly storytelling dances lie between nondramatic and dramatic dance. In the Thai *praleng,* two performers wearing god masks and holding peacock feathers in both hands perform an offertory dance to the god before the main dance-play begins. The Balinese *legong,* danced by a pair of preadolescent girls, may have only the most tenuous dramatic content. Its interest lies in the girls' unison rapid foot movements and fluttering movements of eyes and hands. Dramatic dance is seen at its best in full dance-dramas and in the excerpts from them that are sometimes danced in concert form.

Use of eyes and hands

Styles and conventions of movement and costuming. General characteristics of both dramatic and nondramatic dance are (1) slowness of tempo except in battle scenes, (2) controlled and reserved movements rather than expansive ones, (3) little of the leaping typical of Western ballet but, instead, a feeling of closeness to the ground, and (4) extensive use of arm and hand gestures. From Indian dance has come an open and flexed position of the legs, a side-to-side sliding movement of the head and neck, and a rigidly codified vocabulary of hand and finger gestures known as *mudrās* or *hastās* in India. In most cases the Indian elements have been altered greatly over their 1,000-year period of assimilation. In Thai, Cambodian, and Lao dance, the 24 to 32 Indian *mudrās* have been reduced to nine; in Javanese dance seven can be recognized, and in Bali only one or two. They have also been altered in their shape, and the many specific meanings attached to each in India have become fewer, while in some cases a gesture has no specific meaning. Such hand gestures as shading the eyes and tying the sash, which appear in Javanese dances, are unknown in India. Foot movements in India typically follow the rhythm of a drum, often with vigorous stamping sounds that are emphasized by bells on the ankles, but such movements are virtually absent in Southeast Asia. The exaggerated eye, eyebrow, cheek, mouth, and chin movements through which the Indian dancer expresses a broad gamut of emotions are nowhere to be seen. Balinese dancers use darting eye movements, but the court dancer's face is composed into an almost unchanging expression of aloof gentility.

Close contact between neighbouring countries has led to the development of two regional Indian-influenced dance styles, one for Thailand, Cambodia, Laos, and Burma and one for Indonesia and Malaysia. Characteristics of the former style include the soft *pi phat* music of bamboo xylophones, drums, gongs, and oboe as accompaniment, bent-back finger positions not seen elsewhere in Asia, similar and often identical movements for male and female roles, courtship dances in which lovers touch each other and move in unison, and, in dance-drama, lengthy pure-dance pieces inserted solely for their beauty. In the latter style, the performance is accompanied by music of the gongs and metal bars of the gamelan orchestra. Scarves draped from the waist or neck are flicked for effect and manipulated to indicate strength or flying, and male and female dance are clearly distinguished by the powerful masculine lunges of the men and the tiny steps of the women, who also dexterously manipulate the train of the skirt with their feet. Visually, the mainland dance sparkles. Costumes of brilliant silk are covered with sequins and even jewels, and golden crowns and sparkling body ornaments glitter with reflected light. The male dancer in Indonesia wears a soft batik skirt of brown and white, the female a black velvet bodice. Arms and shoulders are bare and powdered golden brown, creating a subdued and warm effect.

Courtship dances

The main style in Vietnam, apart from folk dance, is dramatic and highly pantomimic, like the movements of Chinese opera. In classical opera, the flowing white sleeves and the pheasant feathers bobbing from the general's headdress are twirled and flicked by the actor in many conventionalized movements derived from Chinese forms. Battle scenes are choreographed into precise dance patterns, but the acrobatic movements common in Chinese opera are seldom seen.

CHARACTERISTICS OF DRAMA

Thematic origins and materials. Most traditional plays and dramatic dances are derived from mythological and legendary sources. The tribal epics that relate the origin of the Ifugao and the Bicolano peoples in the Philippines and a number of animistic stories in Indonesian shadow theatre are indigenous myths of great age, while the widely used, romantic *Pandji* cycle from Java and the Thai *King Abhai Mani* and *Khun Chang Khun Phan* are more recent local legends. The most important dramatic sources, however, are borrowed from the Indian *Rāmāyaṇa* and *Mahābhārata* epics, from the *Jātaka* Buddhist

birth stories, from Chinese novels (such as *The Romance of the Three Kingdoms*) and Chinese operas, and from a host of Islāmic stories, including the *Thousand and One Nights* and the Amīr Ḥamzah tales. These foreign stories are turned into local legends. For example, the Indian Prince Rāma becomes a Thai, a Balinese, or a Javanese prince, embodying the heroic traits admired in each of these countries.

Theme of the restoration of earthly harmony

Plays are invariably extensive and have many scenes. It is not unusual for a play to present action over several generations, an indication of the value placed on cultural continuity. A recurring theme concerns restoration of harmony on earth by a ruler acting in accord with divine law. A kingdom is restored, a prince unjustly exiled returns to assume his throne, a usurper is punished, or the prosperity of the land is assured by consummating a particularly desirable marriage. As in Western drama, the hero gains his ends through struggle. Because he acts as the human representative on earth of the known cosmic will, however, his actions exhibit a natural sweetness and serenity, even in the midst of violence, that is foreign to Western drama. Meditation is often the means whereby the hero gains the power to achieve his goal. In more recent plays based on local history and on contemporary events, the assumption of cosmic harmony has been muted, and emphasis has shifted to depicting human conflicts—nationalist versus Western colonialist, modern daughter versus conservative parents, for example—that may or may not resolve happily.

Characters. Gods, demigods, kings descended from the gods, and princes and princesses are the heroes and heroines of traditional drama and dance. Powerful religious seers advise them, allies and ministers serve them, crude foreign ogres oppose them, and grotesque, slapstick clown-servants are their attendants. The clowns have been the subject of much speculation. Like the *vidūṣaka* clown of Indian Sanskrit drama, they are gluttons, practical and even cynical, and confidants to their masters' passions and weaknesses. Scholars have theorized that the chief Javanese clown figure, Semar, is derived from an ancient Javanese god who was deposed from his supreme position by the introduction into the drama of the later Hindu gods. In the midst of mythological plays, the clowns comment irreverently on political or social issues of the day, seemingly as spokesmen for the common man in an otherwise aristocratic world. Comic and serious scenes alternate.

Improvisation by variation

Dramatic materials. A written script may be used as the starting point for performance, but usually actors, dancers, musicians, and stage crew improvise from a brief scenario. Specific musical selections are matched to certain kinds of scenes, characters, or actions, and standard movements for entrances and exits are known. Standard descriptive phrases of the kind common in all oral literature are used to introduce the hero and his kingdom, and more than a dozen types of recurring scenes are identifiable. A major interest in playgoing lies in perceiving the skill with which performers rearrange and subtly vary these familiar elements from play to play. Narrative commentary accompanying the dances often interprets a specific action in its broad context, thus helping to universalize the theatrical experience.

Costumes, makeup, and settings. Costume and makeup have great importance in plays and dances. By means of elaborate systems of changing the cut, colour, and ornamentation of costume, the shape of the hairdress, the configuration of the crown, or the facial delineation and colour of masks, at least 300 different dance and dramatic characters can be identified. Doll- and shadow-puppet figures are carved according to similarly elaborate means of identification. Persons familiar with a dance or theatrical form can identify most characters by name or by type. Costumes, masks, and puppets may be works of art highly prized in themselves. Court and folk performances once used no scenery at all. Canvas scenery depicting stock scenes is now used by most popular troupes, but unfortunately it is often as inartistic as it is inexpensive. Only the Thai National Theatre, major troupes performing the popular *cai luong* drama in Vietnam, and troupes performing in the Western tradition throughout Southeast Asia attempt to design three-dimensional scenery for each play.

ORIGINS AND DEVELOPMENT OF THE PERFORMING ARTS

Prehistory and links to the present. Knowledge of prehistoric performing arts is necessarily slight. That the performing arts were known and apparently widely practiced by the prehistoric peoples who had settled the mainland and the island archipelagoes is suggested by large bronze drums cast before the time of Christ, numerous pre-Hindu tribal myths in remote areas of the Philippines and elsewhere, masked dances of many types still performed by isolated tribes in Kalimantan (Borneo) and in New Guinea, and descriptions of music and dance by Chinese visitors beginning as early as the 1st century AD. Simple dances were almost certainly accompanied by rhythmic percussion sounds and probably by the tuned metal bars or gongs thought to be indigenous to Southeast Asia. Some scholars suggest that tribal ancestors, animistic spirits, and animals were represented, perhaps in shadow form. Whatever their nature, these were folk performances, in part religious rites connected with seasonal festivals and in part joyful entertainment.

The *trott*, *barong*, and *nat*

A number of existing dances and dramatic forms show prehistoric links. In the *trott*, a Cambodian deer-hunting dance, masked dancers representing hunter, demon, bull, girls, and deer enact the ritual of a deer hunt to ensure its success in real life. The Dayak of Kalimantan perform a dance to exorcise sickness. The *barong* dance-drama of Bali is staged by a village in which malicious spiritual forces are believed to have gained dominance over protective ones. By enacting the stand-off battle between the protective Barong lion figure and the destructive Rangda witch figure, the village ritually restores an equilibrium between the contending forces. A local *nat*, or animistic spirit, of which there are 37 in Burma, can be invoked by the dance of a professional "spirit wife," or *natkadaw*, through whom the *nat* communicates with the living. A disputed theory holds that the shadow play began as a ritual in which the spirits of magically powerful tribal ancestors were called to earth, in their natural form as shadows or shades, for advice.

Spreading of styles. Between *c.* AD 100 and 1000, dance and drama in Southeast Asia were profoundly affected by

Josephine Powell Rome

Apsaras, heavenly dancing girls, bas-relief from Angkor Wat, Angkor, Cambodia, early 12th century.

the introduction of dance style and the vast Hindu historical epics of India. First in Cambodia, then in turn in Thailand, Laos, and Burma, the epic *Rāmāyaṇa* became the source of dance and shadow plays. In Java the *Mahābhārata* dominated, whereas in Bali and Malaysia both epics were popular. Indian influence, however, can be exaggerated. There is no evidence that Sanskrit play texts or written dramatic treatises such as the *Nātya-Śāstra* became known. Strong local performing traditions made it possible to assimilate elements of Indian dance and Hindu stories, and, in subsequent development, Southeast Asian dance and theatre grew ever farther away from Indian styles.

Copper inscriptions from Java identify clowns, actors, musicians, and possibly puppeteers in the 9th century, and epic literature of succeeding centuries contains numerous descriptions of shadow plays that were popular and emotionally gripping. By at least the 4th century, epic recitations were a part of the Brahmanic worship of ancient Cambodia. Carvings of the beautiful *apsaras*, or heavenly dancing girls, adorning the temples of Angkor attest to the importance of court dance in Cambodia between the 10th and 13th centuries.

Accidents of history often carried the performing arts across national boundaries. It is believed King Jayavarman II brought dancers and musicians from Java when he left there in 802 to establish the Khmer dynasty in Cambodia, and shadow puppeteers may have accompanied him as well. Another theory suggests that Cambodia received the shadow play from India by way of Malaysia, through conquest by a Malay prince in 1002. Accidents of war took Khmer dance (and perhaps shadow theatre) first to Laos, when in 1353 a prince who had been raised at Angkor established an independent Lao court at Luang Prabang. Next, it reached the Thai capital at Ayutthaya in 1431, when Angkor fell to invading Thai armies. These returned to their court with the Cambodian court-dance troupe, thereby beginning the traditions of Thai court dance and dance-drama. In 1767 the Thai court was captured, in turn, by the Burmese, who brought to Burma the Thai-modified Khmer dance and created Burmese court drama. By this time, also, Javanese shadow theatre had been taken by colonists to Bali and to Malaysia, from whence it later entered southern Thailand.

Migration of the shadow play

When Indonesia was converted to Islām and Chinese influence became strong in the northern tier of mainland states beginning in the 13th and 14th centuries, existing court dance and dramatic forms were scarcely affected. Instead, new Islāmic plays were devised in Indonesia and Malaysia for shadow presentation and for the doll-puppet theatre. Islāmic influence was very strong in Malaysia, however, and even such pre-Islāmic forms as the shadow play absorbed Islāmic prayers, characters, and themes. Bali was never converted to Islām, and its performing arts are thought to reflect, even today, an older tradition than is seen in Java.

Chinese performing arts came to dominate Vietnam during the 1,000-year rule of northern Vietnam by the Chinese. Long after the Chinese were expelled, Vietnamese kings patterned their dances and opera on Chinese models. In time, however, local Vietnamese melodies and stories took their place alongside those of Chinese origin; and play scripts, at first filled with Chinese loan words, were rewritten in more colloquial Vietnamese.

Popular theatre and Western rule. From the 19th century onward, the incursion of Western culture brought about a variety of developments. A steady decline in the power of the royal courts precipitated the death of court drama in Burma; the shifting of support for dance and drama from the court to national bureaus of education and culture in Thailand, Cambodia, and Vietnam; and the movement of the court dance-drama into the popular theatre tradition in Java. In every country, new popular forms of theatre were created. These were based on historical events, on Islāmic and Chinese stories (but romances rather than Hindu and Buddhist myths), on national heroes fighting colonial rule, and on stories about contemporary events. It was not Western drama that sparked the burgeoning of popular theatre, though these plays were largely spoken dramas interspersed with music and dances. Rather, it was more of an indirect response to colonial rule, which caused an upsurge of nationalist feelings, and to the rapid growth of cities that created large populations without access to either folk or court theatre yet eager for some form of entertainment.

DIVERSE NATIONAL FORMS AND TRADITIONS

Although most of the dance and dramatic forms of Southeast Asia are related at least in the distant past, except in Vietnam and the Philippines, they acquired a very distinctive national and local character over the centuries. An examination of a few of these myriad forms will provide a more precise picture of the dense texture of the performing arts in Southeast Asia.

Cambodia. Court performing arts that had flourished during the Angkor period (802–1431) almost ceased in the centuries following the fall of the Khmer dynasty. Whether there was an organized court life or not is uncertain because of the scarcity of records, but in the 18th and 19th centuries performances in Thai form were produced by the Thai rulers of the western provinces of Cambodia. At Phnom Penh a classical ballet troupe was established by the royal family in the 19th century.

Court styles. The chief court forms are *nang sbek* shadow theatre, *lakon* female dance and dance-drama, and *lakon kawl* male masked pantomime. The puppets of *nang sbek* stand four to five feet in height, have no movable arms, and are manipulated from beneath by two fixed handles or sticks. The standing puppeteer either sways the puppet with his arms or he dances with it. In processional scenes, as many as ten puppeteers parade completely around the screen, front and back. An entire tableau may be carved on one puppet, including several figures, forest scenery, or palace buildings, as if to bring to life the epic scenes carved in relief on the temples of Angkor Wat. Two narrators alternate a slow chant with dialogue. During dance sections, the large *pi phat* ensemble, augmented by a large drum, is played. Only plays based on the *Rāmāyaṇa* are performed, and major puppet figures represent Rāma, his consort Sītā, the monkey Hanumān, and Rāvaṇa, a ten-headed demon king who kidnaps Sītā, Khmer peasant figures have been inserted as rustic clowns in every *nang sbek* play. Performance has religious significance, the gods being invoked and honoured, and a performance may be arranged to assure rain or to halt an epidemic. It is not certain when and how *nang sbek* originated, but it seems probable that it was taken to Thailand in the 15th century and then brought back. This would explain the details of costume and headdress of today's puppets that are in Thai style.

Puppet tableaus

The lithe *apsaras* carved in Angkor's stone show details of the *lakon* style of female dance, but neither these nor other records are evidence that their lively dance was used in relating the epic stories. The 19th-century Thai rulers of western Cambodia reintroduced *lakon* dance and dance-drama, which was indigenous to Thailand as well. At the same time, Thailand's male masked pantomime was brought to Cambodia, as far as is known for the first time, and it became known as *lakon kawl*. Both male and female dance-plays were translated into Cambodian. Recently, costumes and headdresses have been redesigned in the style of the Angkor carvings. The stories, music, dance, and dramatic styles of *lakon* and *lakon kawl* are much like their Thai counterparts.

Popular forms. Lakon bassac, performed by some 20 professional troupes in Cambodia, is a highly eclectic form. Musical selections, dances for female characters, and costuming are borrowed from court *lakon*. The form was created by Khmers living in the Bassac River region of Vietnam. Villains wear Vietnamese costumes and move with Vietnamese opera movements, an evidence of the historical conflicts of the two peoples. Chinese, *Jātaka,* or Khmer stories may be performed. *Pi phat* music alternates with Chinese–Vietnamese instruments and with the Western saxophone and piano. Prince Sihanouk, chief of state between 1941 and 1970, encouraged a few French dramatic productions, but such drama is scarcely known outside the Western-educated elite.

Thailand. Folk *lakon jatri, lakon nai* female dance and

dance-drama, *khon* masked pantomime, and *likay* popular theatre are Thailand's chief performing arts.

Folk performance. *Lakon jatri* began in the south, when male dancer-sorcerers performed, in simple folk style, the *Manora* Buddhist birth story as a dance-play. A troupe of three players was usual. One played the beautiful half-bird, half-human princess, Manora; a second played the hero, Prince Suton; and the third, often masked, played clown, ogre, or animal as needed. Flute, bell cymbal, and drums provided the music. The full *Manora* cycle of plays, staged in a village in the open, could last for two weeks. Probably after the 14th century, some *jatri* troupes moved to the Thai capital, where they established commercial theatres and staged a new all-male drama, *lakon nok* (*nok,* "outside" [the palace]), that emphasized plot and an often obscene humour. Advances in dramatic form were accomplished by court writers of *lakon nok* between 1800 and 1909. *Likay* troupes succeeded and completely supplanted *lakon nok* troupes in the early decades of the 20th century, but such popular *lakon nok* plays as *Sang Thong* ("The Prince of the Golden Conch") are presented today in modified form by the Thai National Theatre.

Bawdy humour

Female court dance-dramas. The *lakon nai* (*nai,* "inside" [the palace]) female dance-drama of the court was created in the mid-18th century from a confluence of three previously separate elements: female court dance, the *lakon nok* drama, and the Javanese *Pandji* stories as subject matter. Romantic episodes from the long *Pandji* tale were ideal for staging in the elegant and delicate style of female court dance, accompanied by songs and the music of a large *pi phat* ensemble. In the unhurried court atmosphere, dance scenes lasted an hour or more, and dance figures might be repeated many times. In time, other stories came to be staged in *lakon nai* and were given other names, but the *Pandji* plays composed by the daughters of King Boromokot (1733–58), by Rama I (1782–1809), and by Rama II (1809–24) remain favourites. In this form, *lakon nai* was introduced into Cambodia within the past two centuries.

Masked mime. Until recent years, a Thai version of the Khmer *nang sbek* shadow play, *nang yai,* occupied an important place in court as a Brahmanic-related ritual performance of the *Rāmāyaṇa.* Thai scholars describe it as the source of *khon* masked pantomime, citing celebrations for King Ramathibodi II in 1515 that included a *nang yai* performance without puppets. Wearing heavy makeup, the puppeteers themselves danced the usual *Rāmāyaṇa* episode as narrators told the story and spoke dialogue. Later, masks took the place of makeup, the screen was eliminated, and *khon* was born. In present-day Cambodia (Kampuchea), one troupe can perform both forms. A number of *lakon nai* elements entered *khon* in later years, so that today a *khon* performance mixes the vigorous, masculine *khon* with gentle *lakon nai* singing style and female dance. All of the Thai dance-drama traditions (*lakon jatri, lakon nok, lakon nai,* and *khon*) are taught at the Department of Fine Arts in Bangkok, and representative plays from them are staged, often mixing traditions, at the Thai National Theatre.

Development of the *khon*

Marie Mattson—Black Star

Rāvaṇa, the demon king, fighting the white monkey Hanumān, in *khon* masked pantomime, Thailand.

Popular plays and puppets. The major popular theatre form is *likay,* which evolved in part out of *lakon nok.* It is now performed by more than 100 troupes in most parts of Thailand. Actors are skilled in improvising not only the dialogue and lyrics but also the plot of a play as well, weaving romantic scenes and fragments of *lakon nai* dance, set to *pi phat* music, into a story from a well-known *Jātaka,* history, or court play. *Likay* plays are set to music of the Lao *khen,* a reed organ, in northeast Thailand. A type of shadow play called *nang talung,* in which a single, seated puppeteer moves small puppets of individual figures with movable arms, is very popular in southern Thailand. The performance technique undoubtedly came from Malaysia, while the plays and the identifying features of the puppet figures, mostly from the *Rāmāyaṇa,* are from Thai *khon* and *lakon nai.* A similar shadow play exists in Cambodia, suggesting that the form travelled from southern Thailand to Cambodia, perhaps in the 19th century.

Laos. From the time Laos became a kingdom in 1353, the performing arts at the relatively small Lao court at Luang Prabang followed those of the more illustrious courts to the south, Angkor in Cambodia and then Ayutthaya and Bangkok in Thailand. Today, Lao dancers study in Bangkok, and the style of dance, music, and drama of the Royal Lao Ballet, the only remaining court troupe in Southeast Asia, is almost identical with that of *lakon nai* in Thailand. It is usual to perform excerpts from the very long dance-plays, the staging of a full-length spectacle being beyond the means of the court at present. Male *khon* dance is known but seldom performed. A number of Lao folk dances are studied and performed by the royal ballet troupe.

Scores of popular troupes perform plays derived from Thai *likay* and set to the lively and melodic Lao folk song style known as *mohlam. Mohlam* balladeers, accompanied by the *khen* (a complex reed organ), have for centuries travelled the Lao-speaking countryside, which includes Laos and northeast Thailand, singing bawdy songs of physical love and weaving into their performance local gossip and bits from the epics and court plays. When *likay* troupes from Bangkok played in northeast Thailand, the *pi phat* music and court dancing were not popular, although the plays themselves were. Enterprising *mohlam* performers then set the *likay* plays to the familiar *mohlam* song style, thereby creating a new popular theatre form, *mohlam luong,* or "story *mohlam.*" Of the *mohlam* troupes, a few large ones are located in major cities in the two countries, but most are small and travel from village to village, performing for a few days or weeks in each.

Balladeers of Laos

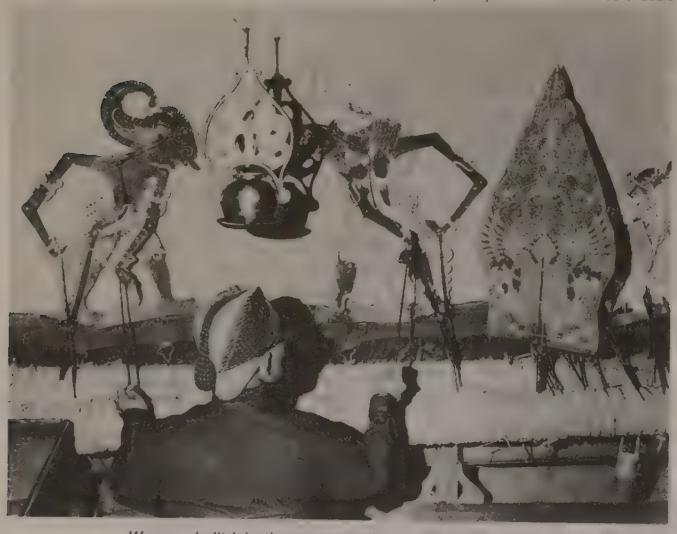
Burma. In spite of an old Burmese tradition of spirit dances stemming from animism and early contact with Indian culture, formal theatre did not begin until 1767, with the introduction of Thai *khon* and *lakon nai* to Burma following the capture and sack of Ayutthaya. Burmese courtiers and dancing girls immediately learned the two forms, and the plays were translated into Burmese. Because Rāma was viewed as a previous incarnation of Buddha, pious Burmese were reluctant to alter *khon* scripts. For a time *Jātaka* plays, including *Rāmāyaṇa* episodes, were forbidden to live actors. Instead, marionette troupes doing plays based on *khon* brought the Rāma stories to the Burmese countryside. But the *Pandji* plays were not considered *Jātakas,* and even the first Burmese version, by U Sa under the title *Inao,* departed from its Thai model, thus setting the stage for the creation of court drama, or *zat pwe,* based on myth and legend but capable of being independently developed. The three *zat* written by U Kyin

U portray the futility of political strife and urge a life of Buddhist renunciation. U Pon Nya created a freer form of dramatic verse, and his *Water Seller* is noted for its comparatively realistic treatment of court life.

Court drama ceased after 1866, when the British conquered Burma. Thereafter, drama was staged by professionals in public theatres, primarily in Rangoon. U Pok Ni in *Konmara* (*c.* 1875), U Ku in *The Orangoutan Brother and Sister* (1875), and others created a new type of drama, *pya zat,* that mixed royalty and commoners, emphasized humour, and added songs to appeal to a popular city audience. Hundreds of these works were published. Popular troupes in Burma today perform a long bill of attractions that lasts most of the night. It comprises songs and dances, a new contemporary play, and, as a final number, a classic *zat* in which remnants of old court music and dance are preserved. British touring companies in the late 19th and early 20th centuries brought examples of contemporary European melodrama and some classics to Burma. Subsequently a number of plays were written in Burmese and in English, following Western conventions and without songs or dance. Of these, *The People Win Through* (1950), by former prime minister U Nu, is among the most interesting examples.

Indonesia. The sober, majestic, and profound court arts of eastern and central Java, where Javanese is spoken, include *wayang kulit* shadow theatre, *wayang orang* unmasked dance, and *wayang topeng* masked dance.

By courtesy of the Indonesian Tourist Board



Wayang kulit (shadow-puppet theatre), Java.

Shadow-puppet theatre. It is uncertain whether the shadow theatre is indigenous to Java or was brought from India, but the *wayang kulit* technique of having a single seated puppeteer who manipulates puppets, sings, chants narration, and speaks dialogue seems to be an Indonesian invention. Unlike most court arts, *wayang kulit* has had centuries of performance in the folk tradition as well, so that today, with several thousand puppeteers active, it is the strongest traditional theatre form in Southeast Asia.

Plays are set in mythological times, some relating to indigenous animistic festivals and worship of local spirits, some directly dramatizing episodes from the *Rāmāyaṇa* and *Mahābhārata* epics, while the majority—the Pandawa (Pāṇḍav in Sanskrit) cycle of about 100 plays—are essentially Javanese creations in which the five heroic Pandawa brothers are placed in different situations. Three and sometimes four god-clown-servants and a set of ogre-antagonists who are not in the epics at all suggest how far removed the shadow plays are from the epics.

The *wayang* puppeteer works within one of the world's most carefully organized performing arts, making possible a virtually solo performance without intermission, from around nine at night until the gray before dawn. Each play is in three parts, coordinated with three keys of music played by the gamelan ensemble. Certain standard scenes appear in a standard order, though some may be dropped. "Opening Audience" introduces the play's conflict, "Inner Palace" shows the king meeting his queen(s), and in "Outer Audience" the army is dispatched. In "Forest Clearing" the first battle scene occurs, and in "Foreign Audience" the antagonist kingdom, usually one of overseas ogres, is introduced. Concluding part one are "Foreign Outer Audience," in which the second army marches forth, and "Opening Skirmish," a battle scene between the two armies. The puppeteer chooses from among 150 musical selections, matched to scene type, character, mood, or action. The puppet figures are carved to indicate character type and status according to fixed patterns for nose, eyes, gaze, stance, body build, and costume. The puppeteer can choose one or another puppet of the same character, coloured gold or black or with a stern or relaxed countenance, to indicate the mood of the figure in a particular scene. In battle scenes, he develops individual encounters between opponents, drawing upon a repertory of 119 movements that are classified for use by god, female, refined hero, muscular hero, ogre, or monkey. Formula narrative phrases describe famous kingdoms and characters, and battles are preceded by challenges couched in standard phrases. Although the puppeteer works only from a brief scenario, he is able to extemporize each performance, adding contemporary jokes for the clowns and molding the performance to suit the occasion and the audience. He and his supporting musicians and female singers are improvising within completely known, although exceptionally complex and subtle, artistic conventions.

Conventions in puppet performances

This artistic system, developed within the shadow theatre for performance of Pandawa plays, has proven to work so well that it has been widely imitated. The entire body of *wayang kulit* drama was adopted in Bali and in Malaysia. At least 25 other play cycles have been performed in Indonesia as shadow drama within this system, including the *Pandji* cycle (*wayang gedog*), Islāmic Amīr Ḥamzah plays (*wayang menak*), and plays dramatizing the revolutionary struggle against the Dutch (*wayang suluh*). The Pandawa *wayang kulit* repertory was transposed to the doll-puppet theatre (*wayang golek*) in Sunda, the western part of Java, and to dance-drama in eastern and central Java (*wayang orang*) and in Bali (*wayang wong*).

Performances are commissioned for special occasions and usually can be interpreted in religious or mystical fashion. There may be offertory plays at harvest time or animistic, ritualistic exorcisms protecting children from being devoured by the voracious god Kala. In *The Reincarnation of Rāma* the divine attributes of the god Wisnu (Vishnu in Sanskrit) reincarnate in Ardjuna (Arjuna), hero of the Pandawa cycle and ancestor of the Javanese race. The translucent screen can be interpreted as heaven, the banana-log stage as earth, the puppets as man, and the puppeteer as god, and the Pandawas can symbolize the manifold attributes of righteous behaviour.

Wayang topeng. Masked dance was also popular at the eastern Javanese courts (*c.* 1000–1400) and may be re-

Masked dance

© Pascal Grellety—Bosviel, Paris

Wayang topeng (masked dance), Bali.

lated to ancient animistic masked dance seen throughout the Pacific islands. Later, Indian dance style was assimilated, and sometime after the 15th century at the earliest, the *Pandji* story was dramatized. This is *wayang topeng*, widely performed as both a sophisticated and a folk art throughout Indonesia. Unlike the large-scale unmasked dance-drama, *topeng* dance focusses on interpreting character through solo dance.

Wayang orang. Java's spectacular dance-drama, *wayang orang*, grew out of the strong unmasked dance tradition that is illustrated in reliefs of female dancers carved on the 9th-century Borobuḍur and Prambanan temples in central Java and that produced the carefully cultivated female group dances of the Surakarta and Jogjakarta courts after their establishment in the 16th century. Of the latter dances, two stand out, the almost sacred *bedaja*, which even today is danced only in court surroundings, and the *srimpi*, in which two pairs of girls execute a delicate slow-motion duel with daggers and bows. In the middle of the 18th century, *wayang kulit's* Rāma and Pandawa plays were set to court dance to form *wayang orang*, or "human" *wayang*. The music, narrative, and dramatic organization of the shadow play was kept largely intact, many of the actors' movements mimicking the stiff actions of the puppets, while new dance sections were added. Court performances stopped with World War II, but *wayang orang* continues to be performed by some 20 to 30 professional troupes in major cities. In popular performances, attractive actresses play the roles of such refined heroes as Ardjuna and humour and spectacle take precedence over dance.

Actresses as refined heroes

Ketoprak and ludruk. Two other types of popular theatre, *ketoprak* and *ludruk*, were performed in Java by 150 to 200 professional troupes. *Ketoprak*, created by a Surakarta court official in 1914, evolved into a spoken drama of Javanese and Islāmic history in which the clown figure is a spokesman of the common man. Whereas *ketoprak* is performed primarily in central Java, *ludruk*, a spoken drama that handles mainly contemporary subject matter, is performed in eastern Java by both amateur and professional troupes. Though *ludruk* is relatively realistic, male actors play all roles. Songs and dances, accompanied by gamelan music, are performed between acts in both forms.

Sundanese performing arts. There are three main performing arts in the Sundanese area of western Java. *Reog*, a kind of urban folk performance, can be seen especially in the streets of Jakarta: two or three men improvise popular songs, dances, and dramatic sketches for a neighbourhood audience in this type of entertainment. *Wayang golek* is a performance based on *wayang kulit* but using doll puppets without a screen. Approximately 500 Sundanese puppeteers perform *wayang golek*. Female singers, who are almost as important as the puppeteer, respond to requests and gifts of money by singing song after song and virtually stopping the play. *Sandiwara* troupes in Jakarta, Bandung, and a score of other cities perform both *wayang* stories in the form of Sundanese dance-drama and spoken historical and contemporary dramas for popular audiences. Sundanese-style court dances and *topeng* masked dances are often performed solo at festivals and for circumcision or wedding celebrations in private homes. Sundanese dance is more sensuous than Javanese and broader in style.

Balinese dance-drama. Of the many factors that have contributed to the remarkable flourishing of dance and drama on the island of Bali for more than a millennium, three are of particular note. First, Bali remained isolated from both Islām and the West. Second, there was a merging of folk and court performance styles into a single communal tradition appreciated by all. Third, dances and plays are indissolubly linked to the recurring cycles of local festivals and rituals whereby the well-being of the community is maintained against constantly threatening malicious forces in the spirit world. From the verve and brilliance of Balinese performances it is clear not only that the people like to perform but also that there exists some culturally determined compulsion to do so.

Balinese dance and dramatic forms are so numerous that only a few can be noted. Balinese villagers playing in the *barong* exorcism dance-drama are not merely actors exercising theatrical skills. The actors' bodies, going into a trance, are believed to receive the spirits of Rangda and the Barong, and it is the spirits themselves that do battle. Thus the performance is actually more a ritual than a piece of theatre. The *sanghyang* dance is usually performed by two young girls who gradually go into a state of trance as women sing in chorus and incense is wafted about them. Supposedly entered by the spirit of the nymph Supraba, the girls rise and dance, often acrobatically, though they have been chosen from among girls untrained in dance. The dance's purpose is to entice Supraba to the village to gain her blessing when evil forces threaten. In the *ketjak*, or monkey dance, as many as 150 village men, sitting

Hypnotic possession of Balinese dancers

Tor Eigeland—Black Star

Ketjak, or monkey dance, Bali.

in concentric circles around a flaming lamp, chant and gesticulate in unison until, in trance, they appear to have become ecstatically possessed by the spirits of monkeys. This performance, however, has no ritual function of altering an earthly condition.

That the Balinese *wayang kulit* may represent the older style of *wayang*, known on Java before the coming of Islām, is suggested by the less stylized shape of the puppets, by the shorter performing time of four to five hours, and by the simple music of only four *gender*, a bronze instrument similar to a xylophone with resonance chambers underneath, from the gamelan ensemble. In one type of shadow play having a special religious significance, the puppets perform before a screen during the daytime, and the puppeteer is seen in his role as a Brahman priest, bare to the waist. In the *redjang* processional dance, village women symbolically offer their bodies to their temple gods.

Because Balinese performing arts are vitally alive, they change from decade to decade, even from year to year. The *gambuh*, respected for its age, contains elements of dramatic dance, song, narrative, and characterization found in later forms. It is thought dull, however, and is seldom performed, though it is believed to have provided the model for the singing style of popular *ardja* opera troupes and the dance style of the lovely girls' *legong*. *Wayang wong* is analogous to the Javanese *wayang orang*, but masks are worn and the repertory is limited to Rāma plays. Pandawa plays are staged in identical style but are called *parwa*. It has been suggested that these forms also stem, at least in part, from *gambuh*. *Wayang topeng* masked-dance plays are ancient, being mentioned in a palm-leaf document of 1058. The Javanese chronicle of the Majapahit period (*c.* 1293–1520), the *Pararaton*, in

which Ken Angrok is the hero, is a favourite *tapeng* story. This points to the strong influence exerted by Javanese on Balinese arts after the Majapahit court was transferred to Bali in the 16th century to escape Islāmic domination.

Malaysia. The Malay peninsula, in the geographical centre of Southeast Asia, has assimilated repeated intrusions of neighbouring cultures. The dances of the former princely states on the east coast show the influence of Indian nondramatic dance.

The multiform wayang. Rulers from Java in the 13th and 14th centuries and later large colonies of Javanese introduced their *wayang kulit* shadow theatre. The puppets of *wayang Djawa,* or "Javanese" *wayang,* are identical with the two-armed, long-nosed, highly stylized puppets of today's Javanese *wayang kulit.* Those of *wayang Melayu,* or "Malayan" *wayang,* have only a single movable arm and are less sophisticated in conception, which suggests that they are either descended from old Javanese puppets, before both arms were made movable, or are a degeneration of the more complex form. Rāma, Pandawa, and Pandji plays are staged. The puppets of *wayang Siam,* or "Siamese" *wayang,* though manipulated by a single seated puppeteer, represent a Thai conception of the figures from the *Rāmāyaṇa;* and costumes, headdresses, ornamentation, and facial features follow those of *khon.* The plays include Islāmic elements as well, while the chief clown figure, Pak Dogol, is thought to be a recent Malay creation that has supplanted Semar, the Javanese clown of *wayang kulit.*

In a performance, puppets of all types may appear together. Either such Thai instruments as the *lakon jatri* drum and small bell cymbals or gamelan instruments play the accompanying music. Song lyrics can be in ancient Javanese; animistic, Islāmic, and Hindu-derived invocations to the gods are offered in the Thai and Malay languages; and the play proper is in colloquial Malay. Puppeteers once performed throughout the peninsula, including the five Malay-speaking provinces of southern Thailand, but today puppeteers are found primarily in northeast Malaysia.

Multilingual performances

Chinese and popular entertainments. Chinese immigrants introduced various forms of opera during the 19th century. Troupes perform for Chinese Buddhist temple festivals, for local fairs, or on national holidays. In Singapore troupes occasionally perform in public theatres as well. Young people of Chinese descent in both Malaysia and Singapore have little interest in the opera, however, because their Chinese is limited. Occasionally troupes import star performers from Hong Kong or tour Chinese communities in Thailand.

Bangsawan was created by professional Malay-speaking actors in the 1920s as light, popular entertainment. Songs and contemporary dances were added to a repertory of dramatic pieces drawn from Islāmic romances and adventure stories. Troupes travelled to Sumatra, Kalimantan, Sunda, and Java, where their melodramatic plays found large audiences and influenced local performers of *sandiwara, ketoprak,* and *ludruk.* The cinema and television, however, have captured much of this audience.

Vietnam. An indication of the antiquity of the performing arts in Vietnam is a large bronze drum of the 3rd century BC found near Haiphong, in northern Vietnam, which is ornamented with instruments and musicians playing for dancers. Chinese performing arts presumably were a part of court life in northern Vietnam during the period of Chinese rule (111 BC–AD 939), and between the 10th and 13th centuries the dances and music of the Hinduized Cham peoples, living in what is now central Vietnam, were welcomed there. The melancholy Cham songs were particularly popular, and most authorities believe that the sad southern style of Vietnamese singing is derived from them.

Satirical drama. Hat cheo is a popular, satirical folk play of northern Vietnam that combines folk songs and dances with humorous sketches criticizing the people's rulers. Some scholars theorize that it is an indigenous folk art, whereas others, to show that it reached the people from the court, cite the legend of a Chinese actor who in 1005 was hired by the Vietnamese king to teach "Chinese satirical theatre" to his courtiers. *Hat cheo* is widely encouraged by the government.

The opera. The classic opera, known as *hat boi, hat bo,* or *hat tuong,* is a Vietnamese adaptation of the Chinese opera long supported by kings and provincial mandarins as a court art and performed for popular audiences as well, especially in central Vietnam. The introduction of Chinese opera is attributed to the capture of a troupe of performers attached to the Mongol army that invaded northern Vietnam in 1285. The actors' lives were spared in return for teaching their art to the Vietnamese. In 1350 another Chinese performer was engaged by the northern court as an instructor. Almost exclusively a court art in the north, *hat boi* was made a form of popular entertainment in central Vietnam by the playwright Dao Duy Tu in the 16th century. It was introduced to southern Vietnam under the Nguyen dynasty in the 18th and 19th centuries, but its future was jeopardized by the decades of war in the mid-20th century. The last large troupe of court musicians, dancers, and actors at Hue in southern Vietnam disbanded in 1945. The postwar government of the late 20th century did not provide *hat boi* with strong support, and the popular troupes lacked audiences.

Blighting effects of war

In form and content, *hat boi* is a blend of China and Vietnam. Direct imitation of Chinese costume and acting techniques was encouraged under the reign (1847–83) of Emperor Tu Duc and it is probable that the present form of *hat boi* dates from this period. At Tu Duc's court in Hue, the playwright and scholar Dao Tan gathered 300 actors and with them wrote out texts of the standard repertory that previously had been preserved orally. He then had the texts published and distributed them to actors and troupe managers. In the 20th century there has been a movement to loosen the rigid structure of *hat boi* and to reduce the high proportion of Chinese loan words that makes the operas difficult for the ordinary Vietnamese to appreciate.

Following Chinese practice, the operas are classified as military or domestic. The former, which may be derived from Chinese and Vietnamese legend or history or may be purely fictional, concern struggles for power between kings. The Chinese novel *The Romance of the Three Kingdoms* furnishes material for many military plays. The latter, dealing with the lives of commoners, contain humorous scenes alternating with scenes of suffering that are played to the accompaniment of sad southern-style songs. The Confucian ethic of obligation to one's superior—of wife to husband, of son to father, or of subject to king—underlies plays of both types.

Hat boi staging is modelled on conventions of Chinese opera. Actors perform on a stage that is bare except for a table and two chairs. These can serve as a castle, a cave, or a bed as well as for sitting and eating. A single embroidered drop at the rear has an entrance right and an exit left. Costume and makeup indicate character type: black for boldness, red for anger or rashness, white for treachery, and gold as the colour of the gods. Conventionalized mime may be used alone or in conjunction with symbolic properties. The actor mimes stepping over an imaginary threshold or sewing without needle and thread, but he indicates riding a horse by gestures with a riding crop and travels in a carriage when a stage assistant holds flags with wheels painted on them at each side of his body. Percussion instruments accompany stage action, and songs—which may be in falsetto Chinese style, in soft southern Vietnamese style, or in a form of prose recitative—are accompanied by stringed instruments.

The popular stage. Southern-style singing is the basis of another type of theatre, *cai luong,* begun in the 1920s by popular singers who performed plays in which they sang the love lament *Vong Co.* Today, regardless of whether a historical or contemporary play is being performed as *cai luong* or which of many troupes is staging it, this melody will be heard throughout the play many times, underlying different lyrics. *Cai luong* stars are lionized, and the best troupes maintain high artistic standards. Among popular theatre forms in Southeast Asia, only *cai luong* plays are fully scripted and directed as they would be in the Western theatre. In contrast to the operetta form of *cai luong,* modern spoken drama is known as *kich.* It is a young dramatic form performed mostly by amateurs who are trying to put Western dramatic conventions into practice.

The love lament

The Philippines. Whatever indigenous theatrical forms may have existed in the Philippines, other than tribal epic recitations, were obliterated by the Spanish to facilitate the spread of Christianity.

The comedia. The earliest known form of organized theatre is the *comedia*, or *moro-moro*, created by Spanish priests. In 1637 a play was written to dramatize the recent capture by a Christian Filipino army of an Islāmic stronghold. It was so popular that other plays were written and staged as folk dramas in Christianized villages throughout the Philippines. All told similar stories of Christian armies defeating the hated Moors. With the decline of Spanish influence, the *comedia*, too, declined in popularity. Some professional troupes performed *comedia* in Manila and provincial capitals prior to World War II. Today it can still be seen at a number of church festivals in villages, where it remains a major social and religious event of the year. Much in the manner of the medieval European mystery-play performances, hundreds of local people donate time and money over several months to mount an impressive performance.

By courtesy of Philippine Embassy

Comedia, or *moro-moro*, folk drama based on the battles between the Christians and the Moros, the Philippines.

Styles from Europe. Dances and dramas from Spain were brought in, some of which took root. The "María Clara," a stately minuet, and the "Rigodón de Honor," a quadrille, were adopted by local European society for its formal balls. Spain's sprightly operetta, the *zarzuela*, became the favourite light entertainment in Manila and other cities. Professional *zarzuela* troupes continued to flourish in the early decades of the 20th century but had disappeared by World War II. New plays with original music were produced in profusion. A number of them based on topical themes and criticizing American colonial policies were banned.

Western drama is studied and widely performed in both English and Tagalog. There are no professional companies, but amateur university and community groups abound. Western classics and recent popular successes are staged, and in recent years many original plays have been written to celebrate the Filipino heritage.

(J.R.B.)

Visual arts

GENERAL CONSIDERATIONS

Religious–aesthetic traditions. The visual arts in Southeast Asia have followed two major traditions.

Indigenous magical and animist tradition. The first is a complex inheritance of magical and animist art shared by the different tribal peoples of the mainland, where it evolved from Paleolithic origins, and of the islands. Such art gave the peoples who made it a sense of their identity in relation to the forces of their natural environment, to the structure of their society, and to time. It consists of types of potent emblem, mask, and ancestral figures broadly similar to those that hunters and early farmers the world over have used in connection with seasonal ceremonies, life and death rituals, and ecstatic shamanism (belief in an unseen world of gods, demons, and ancestral spirits responsive only to the shamans, or priests). The spiritual powers that the arts name and invoke are local and vary from group to group of the population. The rich formal artistic languages have been subject to successive episodes of influence from inland Asia, but each of the tribal groupings has developed its own artistic language on the basis of a common fund of Southeast Asian thought forms.

Indian tradition. The second major tradition was received from India during the early centuries of the Christian Era, when seagoing merchants from that subcontinent so fertile in ideas were expanding their trading activity. Into many parts of Southeast Asia—especially Burma, Thailand, and the coasts of Cambodia and Indonesia, where Indian traders settled and married into the families of local chieftains—they brought with them a script and literature in the sophisticated Sanskrit language. They also brought a highly developed conceptual system dealing with kingship, statecraft, and hydraulic engineering, integrated and authenticated by profound metaphysical ideologies of Indian pattern, both Hindu and Buddhist. These ideologies claimed to be universal, embracing all human diversity within a cosmic frame of reference. And this explains why the culture was adopted. For there was no Indian conquest of terrain; instead, the Indian conceptions, along with the art that expressed them, were used by dynasties in the colonial kingdoms as a method of overcoming divisions in their population, of centralizing effort, and of uniting their religions into viable states based upon cities. Although the new religious conceptions must have offered deep personal satisfaction to the general population, the architecture and sculpture in stone and bronze in which they were artistically expressed were expensive in materials, labour, and skill and were thus available primarily to patrons who were claiming for themselves a royal (*i.e.*, divine) status and using the resources of art to demonstrate that status.

Why the Indian culture was adopted in Southeast Asia

The Indianizing traditions were continually refreshed by direct influences from India and Ceylon. There can be very little doubt that, during the early centuries after the birth of Christ, Indian artists and craftsmen travelled to work in the distant trading colonies of Southeast Asia, for they would have been needed to set up local traditions with proper formulae and methods. And there can be no doubt either that works of art made in India were continuously exported to the colonial kingdoms, thus keeping the local art styles in touch with developments "at home." It is also clear, however, that within a very short space of time the Southeast Asian kingdoms produced their own distinctive local versions of Indian styles; and some of their work shows skill, finesse, and invention on a colossal scale unrivalled even in India.

Although the art styles were to some extent sectarian, and sectarian partisanship played a part in political events, it was by no means unusual to find Hinduism and different forms of Buddhism flourishing side by side. In both Burma and Thailand, however, dynastic options were early exercised in favour of that particular form of Buddhism known as Theravāda (Hīnayāna), which adheres to the nontheistic ideal of purification of the self to Nirvāṇa. These countries follow the same form to the present day. It was also adopted in Cambodia and southern Vietnam after prolonged and successful periods of Hindu and Mahāyāna (a theistic branch teaching compassion and universal salvation) Buddhist dominance. The strongly Sinicized population of the region around the Gulf of Tonkin, which pushed gradually down the coast of Vietnam to become the modern plains Vietnamese, began to adopt Theravāda Buddhism with its artistic types by about the 13th century AD, partly because this form could be best adapted to its self-contained and antidynastic cellular social structure.

Theravāda and Mahāyāna Buddhism

Relations between the two traditions. Even in those regions where Indian influence became strongly entrenched, the older layers of more primitive religion and artistic consciousness remained very much alive. Indian deities were readily identified with local spirits. The tribal populations retained, as many still do, their old animist customs, especially those connected with fertility and practical magic, often with an art (in perishable materials) in which to express them. These arts were influenced by and exercised a reciprocal influence upon the styles of officially imposed Indianized arts. In many parts of Southeast Asia, where the official Indian styles were not completely established (most of Borneo) or where they died out (colonies in Celebes), in inaccessible areas beyond the reach of dynastic influence, or on isolated islands, primitive styles have survived unmodified. Even in Indianized regions where a strict formula, say, for a necessary building type, had not been imported, a native pattern was adopted into the official canon (*e.g.*, Laos). In the Indonesian island of Bali, which has remained nominally Hindu, the Indian and the folk elements were thoroughly assimilated to each other, producing a quite individual style of both religion and art. In Sumatra and Java, whose populations were gradually converted to Islām from India during the 13th–16th centuries, the primitive cult of the ancestors was revived and encouraged by Muslim rulers, with folk versions of denatured Hindu art adapted to it. Decorative styles based on this art have flourished there and have been officially revivified during recent years. In the Philippines, notably in and around Manila, Spanish Roman Catholic art flourished after the Spanish colonization of 1571.

The royal temple

Artistic styles. The royal temple is the basis for the classic Indianizing styles of Southeast Asia. Each Hindu temple is centred on a shrine, symbolizing heaven upon earth, which is crowned by a roof tower representing the cosmic Indian mountain, Meru, conceived as the hub of creation. Since all the peoples of Southeast Asia already believed the natural habitat of spirits and gods to be a mountaintop, the Indian pattern was readily accepted. The temple usually stands upon a lofty terraced plinth (a block serving as a base), which itself also symbolized a mountain. Towered shrines could be multiplied on the terraces, though one of them remained the principal focus. Within the cell of this main shrine was a sacred image carved in stone or cast in bronze. The local Hindu ruler identified the subject of this image as his transcendent patron, or celestial alter ego. This was normally one of the Indian high gods, Śiva (represented perhaps by a phallic emblem, the *liṅga*) or Vishnu. In Mahāyāna Buddhist kingdoms a royal *bodhisattva* (a being that refrains from entering Nirvāṇa in order to save others) was sometimes adopted to fulfill the same role, a favourite form being known as Lokanātha, or Lokeśvara, Lord of the World. Subsidiary shrines, niches, or terraces sometimes contained subsidiary images, including goddesses representing at the same time wives of the god and queens of the king. These images were worked in smooth, deeply rounded, and sensuously emphatic styles derived from Indian art but with varying inflections characteristic of each region and time. The whole exterior of the shrine was usually adorned with rhythmic moldings, foliage, and scrollwork, with figures representing the inhabitants of the heavens. Ideally, the building was constructed and carved in stone; but, particularly where good stone was not readily available (for example, in Burmese Pagan), it could also be brick, coated and sculptured with stucco after northeast Indian patterns. Temple complexes tended to grow as successive kings strove to outdo their predecessors with the magnificence of their buildings. Hindu rulers, influenced perhaps by vestiges of tribal custom, would sometimes retain their own family's temples and images while destroying those of earlier dynasties.

Buddhism, however, is a religion based on a doctrine of transcendent merit and sustained by an order of monks who have, ultimately, no vested interest in kings and gods. They may, however, take a great interest in the world of spirits and the operations of astrology, just as the local population does, even though they regard such matters as subordinate to the ultimate Buddhist aim of universal Nirvāṇa. Buddhist monasteries, therefore, tended to expand around stupas (domed monuments emblematic of the Buddhist truth, also called pagodas or *dāgabas*) of ever-increasing size and number; the preaching halls, libraries, and living quarters for monks were continually enlarged and repeatedly rebuilt, often as a testimony to the piety of royal patrons. Although, strictly speaking, Theravāda Buddhism has no place for a "divine ruler" whose identity an actual king may adopt, provision was made in legend and in court and monastic ritual for the ruler of a Theravāda country to assume a magical role as the dominant sponsor and patron of the Buddhist truth. His legendary prototype was usually not identified, therefore, with an icon of the enlightened Buddha but with images such as the chief disciple at the knee of the enlightened Buddha, as Prince Siddhārtha (the Buddha-to-be), or figuring in scenes of the Buddha's life that lined the monastery halls and corridors.

Prescriptive formulas for art

Both Hindu and Buddhist art were produced according to prescriptive formulas. If the formulas were not followed, the art was believed not to fulfill its transcendent function. In practice, however, there has been room for styles and types of image to change and develop fairly quickly. Hindu and non-Theravāda art recognized what could be called aesthetic values as a component in religious expression. Theravāda Buddhism, however, which might be called fundamentalist, has always attempted to preserve the closest possible connections with the Buddha's recorded original deeds and sayings; its art, therefore, has concentrated on repeating in its main Buddha figures the most exact possible imitations of authentic ancient images. This had led to a relative monotony of style in Theravāda icons (see below *Burma; Thailand and Laos*). In the subsidiary sculptured and painted figures, however, which illustrate scenes from sacred history, Theravāda art has had greater freedom of invention. In the 20th century, Theravāda Buddhism is the only form of Indian religion to survive in Southeast Asia, save for the modified Hinduism of Bali. Its architecture in recent centuries has been decorated with a vigorous, sometimes coarse fantasy and made gaudy with gilt paint and coloured glass.

General development of Southeast Asian art. Most of the works made under the inspiration of the primitive, magical, and animist tradition are in perishable materials such as wood. Because the climate is so hostile, the works that survive are relatively recent; and any that is even 100 years old generally owes its preservation to Western interest. There are, however, a large number of Neolithic stone implements and prehistoric stone monuments (megaliths), as well as bronzes, which provide a solid archaeological basis for interpretation of Southeast Asian primitive art.

For the art of the classic Indianizing civilizations, French archaeology played the major role in clearing, excavating, and reconstructing major sites in Cambodia, Laos, and Vietnam; Dutch archaeology in Indonesia; British in Burma. Old bronzes have been found in fair quantities; apart from those of the early Dong Son culture (see below *Bronze Age: Dong Son culture*), all belong to one or other of the Indianizing traditions. Many old brick and stucco buildings survive, notably the medieval work at Pagan and in central Thailand, though an enormous number are known to have perished. Little very old painting is known, save a few Indianizing medieval rock and wall paintings on plaster. In spite of the fact that Buddhist monasteries are able to act as agents for preserving their own artworks, most of the surviving Buddhist pictorial art on wooden panels or other fragile material is less than 300 years old.

The stone of dynastic buildings, of course, has survived far the best. Scholars thus know much more about Indianizing stone architecture, with its sculpture, than about any other Southeast Asian visual art. But, where good relief sculpture flourished, one can legitimately assume that vanished pictorial arts also flourished; and from details carved in stone and incised on bronze, as well as from the scattered enthusiastic references in Chinese sources, one can be sure that throughout their history the Southeast Asian peoples have been intensely creative and have lived their lives surrounded by a wealth of imaginative art in many different mediums.

Unexcavated sites

There are many sites yet to be discovered and excavated. Twentieth-century knowledge of the history of art in many parts of Southeast Asia, especially of important episodes in Burma, Thailand, and Sumatra, is still scantily documented.

Neolithic period. The earliest works in Southeast Asia that can be called art are the rectangular polished axheads of a familiar late Neolithic type that have been found at many sites in Malaya, Indochina, and Indonesia. Some of the later Neolithic (*c.* 2000 BC to early centuries AD) implements are extremely beautiful and polished with the greatest care. They include practical adzes and axes; but some, made of semiprecious stone, were clearly intended for purely ritual purposes. Even in the 20th century a few such blades are preserved and revered as sacred objects in certain Indonesian farming communities and similar objects have continued to be made in some very remote regions. These tools, with their fine edges, suggest that their owners were capable of very high quality woodworking and might well have decorated their wooden houses with designs of which we know nothing.

During the Neolithic Period, metal—both bronze and iron—came into use for implements but did not greatly alter the material culture. In many regions, notably Cambodia, Borneo, and Sumatra, numerous works of megalithic, or stone, art survive, including menhirs (single upright monoliths), dolmens (two or more upright monoliths supporting a horizontal slab), cist graves (Neolithic graves lined with stone slabs), and terraced burial mounds, all dating from the late Neolithic epoch. Some remarkable large stones are worked in relief with symbols and with images of animals and men, notably in the Pasemah region of Sumatra. Shaped stone sarcophagi and skull troughs (containers to hold the skulls of ancestors and of enemies at village shrines) are also known. These megalithic art objects suggest a highly developed cult of a spirit world connected with the remains of the dead (see below *Cambodia and Vietnam; Indonesia*).

Bronze age: Dong Son culture (c. 4th–1st centuries BC). By about 300 BC a civilization with elaborate arts based on bronzeworking existed, extending probably from the Tongking region into Laos, Vietnam, Cambodia, and Indonesia. This is called for convenience, after a major site, the Dong Son culture, though it may not have been a true cultural unity. A variety of bronze ritual works, many decorated with human and animal figures and with masks, were cast by the cire perdue method (metal casting using a wax model). The chief objects were ceremonial drums, large and small; the largest was found in Bali and is called "the Moon of Bali" (see below *Indonesia*). Extremely elaborate bronze ceremonial axes were made—probably as emblems of power. Certain relief patterns on the bronzes suggest that "ship of the dead" designs, like those still woven in textiles in both Borneo and Sumatra, may well have been woven even then. The spiral is a frequent Dong Son decorative motif; later Dong Son art was probably responsible for transmitting—especially into Vietnam, Cambodia, and Borneo—versions of the contemporary Chinese Chou dynasty's asymmetrical squared-hook patterns.

1st to 10th century. During the 1st century AD, Indian influence began to spread through Southeast Asia in the wake of trade, both overland, through Burma and Thailand, and by sea traders settled at especially favourable spots along the inland roads and along the sea routes around the coasts and into the islands. Buddhism, which was particularly popular among the Indian merchant classes, took root at a large number of trading cities, where monasteries were set up under the patronage of local kings. Many fragmentary Buddha images based upon Indian types of *c.* AD 300–400 have been found in Burma, Thailand, and Cambodia, produced in the kingdoms of the Mon people, the chief of which, in Thailand, was called Dvaravati. By the 5th century the first Hindu kingdoms had been established in western Java and Borneo. These kingdoms produced dynastic cult images, fragments of which have been found.

The first Hindu kingdoms

Perhaps the most splendid of the earlier Indianizing kingdoms, lasting till the 9th century AD, was that of the Pyu people in the upper Irrawaddy River Valley. The Pyu were the people most directly in touch with eastern India by land routes. Only one of their enormous cities has been explored archaeologically (see below *Burma*). The remains of Buddhist buildings, east Indian Buddhist images, and Hindu sculptures of Vishnu have been found there.

In the 1st century AD the predominantly Hindu kingdom known as Funan (the name given it by Chinese historians) was established in Cambodia. It seems to have controlled an empire that included kingdoms in Malaya and even parts of southern Burma. Its population was probably Mon and shared the culture of the Mon in the lower Irrawaddy Basin. (The Funan kingdom really represents the earliest phase of what became, in the 9th century, the great Cambodian Khmer Empire.) Between *c.* 550 and 680 the kingdom retreated from the coast up the Mekong River into Laos, where it was called by the Chinese Chenla. This joint Funan–Chenla tradition produced some of the world's most magnificent stone cult images. Though Buddhist icons are known, these images principally represent Hindu deities including Vishnu, his incarnation Krishna, Śiva, and a combined Śiva-Vishnu figure called Harihara. The images were housed in wooden or brick shrines, now vanished.

During the Chenla retreat the Theravāda Buddhist kingdom of Dvaravati flourished in southern Thailand, on the lower reaches of the Mae Nam Chao Phraya; the kingdom lasted until the 11th century, when it was captured by the Khmer. What little of its art is known is close to that of eastern India and provided the basis for later Buddhist art in the Khmer Empire, as well as for some of the later forms of Thai art.

The central Javanese kingdom

Almost contemporary with Chenla was the rise of the central Javanese kingdom. Soon after AD 600 the earliest surviving Hindu temples were built. In *c.* 770 the Śailendra dynasty began its long series of superb stonecut monuments both Hindu and Buddhist, which culminate in two enormous symbolic architectural complexes: the Mahāyāna Buddhist Borobuḍur (*c.* 800) and the Hindu Lara Jonggrang, at Prambanam (*c.* 900–930). These monuments were decorated in an individual and exceptionally accomplished style of full-round and relief sculpture. Many small bronze religious images have survived. The art of the Śailendra dynasty testifies to the imperial and maritime power of the central Javanese kingdom, which seems to have influenced politics and art in Khmer Cambodia. It also took over the possessions of a major Theravāda Buddhist kingdom called Śrivijaya, which had flourished in Malaya and Sumatra and was centred at Palembang. The Javanese Śailendra ruled most of Malaya and Sumatra and installed themselves there in the mid-9th century, when their home terrain in Java was taken over by the Mataram dynasty, heralding the eastern Javanese period, which began in 927. Śrivijaya, under Śailendra rule, declined in the mid-11th century, and most of its remains still await discovery.

In Vietnam *c.* the 2nd century AD the predominantly Hindu kingdom of Champa was founded. Its capital was at My Son, where many temples have been found. This kingdom suffered much from attacks by the Chinese, and, after it began to lose the north to the Sinicized Vietnamese in 1069, the Cham capital moved in 1069 to Vijaya (Binh Dinh), in the south. There it was involved in continual warfare with the Khmer, who finally annexed southern Vietnam in 1203. The art of the northern Vietnamese as a whole has always been so strongly under the influence of China that it can best be characterized as a provincial Chinese style.

The Khmer Empire

10th century to the present. In Cambodia the Khmer Empire succeeded to the old territories of Funan–Chenla. About 790 the first major Khmer ruler, Jayavarman II, who was related to the old Funan royal family, came to Cambodia from the Śailendra court in Java. In 802 he set up a religious capital on a hill at Phnom Kulen; he seems to have called in artists from Champa and Java, thus giving to Khmer art a distinct new impetus. At another site, Sambhupura (Sambor), he built temples with sculpture based upon the old Funan–Chenla tradition. At Amarendrapura, about 800, he built a brick pyramid—an artificial mountain—to support a quincunx of temples.

It was Indravarman I (877–889) who laid the foundations of the fabulous temple complex known as Angkor. His plan was based on a rectangular grid of reservoirs, canals, and irrigation channels to control the waters of the river system. Later kings elaborated this original design to a colossal scale. Indravarman built the first great works of Khmer architecture: the Preah Ko, at Roluos, and at Angkor his temple mountain, the Bakong, ornamented with sculpture. Successive kings built their own temple mountains there, including the Bakheng (*c.* 893), Pre Rup (*c.* 961), the Ta Keo (*c.* 1000), and the Baphuon (*c.* 1050–66) and culminating in Angkor Wat, built in the first half of the 12th century by Suryavarman II. After a disastrous invasion by the Cham, Jayavarman VII undertook the most ambitious scheme of all, the Mahāyāna Buddhist Angkor Thom and Bayon (*c.* 1200). Thereafter, for a variety of reasons, including conquest by the Thai, no more large-scale work was done by Angkor and the country became Theravāda Buddhist. The modern dynasty has adapted remnants of traditional splendour, and the craftsmen of Cambodia have remained capable of work in the same vein as but often superior to the Thai.

Hindu Javanese art continued to be made under the eastern Javanese dynasties (1222–14th century), although their structures are not nearly as ambitious as the central Javanese works. There are many temple enclosures and volcanic bathing places with modest stonecut architecture. Some of the stone sculptures from these sites, however, are now world famous. In the 20th century the east Javanese tradition still survives, modified by folk elements, in Bali, to which the east Javanese Hindu kings retreated in the 16th century to maintain their religious independence in the face of Muslim expansion. Muslim monuments in the form of mosques and tombs are found in various parts of Indonesia. They adapt older forms of Indonesian art.

In 1056 the great Burmese king Anawrahta decreed Theravāda Buddhism to be the religion of his country, replacing earlier cults. He removed the Mon monks and artists from the capital of the old Mon kingdom in southern Burma, transporting them to his own northern capital, Pagan. There they built a city, with many large brick and stucco temples (pagodas) based on Indian patterns, that remains one of the most impressive sites in Asia. The Mongol invasion of 1287 put a stop to work there.

The Mon kingdom, Dvaravati, of southern Thailand, was annexed to the Khmer Empire in the 11th century and Khmer imperial shrines were built there. After the decline of the Khmer and the Mongol invasion of 1287, a powerful alliance of racially Thai kings established the first major Thai empire, retaining Theravāda Buddhism as the state religion. Thailand was divided into two principal regions, northern and southern, with capitals, respectively, at Chiang Mai and Ayutthaya, possession of the trade city of Sukhothai being an issue between them. In all the Thai cities, brick and stucco temples were built on variants of Indian and Burmese patterns. Many fine bronze Buddha figures, large and small, were cast in canonical Theravāda Buddhist styles. Most of these figures were accommodated in monastery halls built in impermanent materials.

In both Burma and Thailand a very large number of monasteries, usually surrounding one or two principal pagodas, were constructed during the later Middle Ages and into modern times. The major cities of Rangoon, Mandalay, and Bangkok contain the most elaborate examples, although there are many elsewhere. Because the pagodas were repeatedly enlarged and redecorated and the wooden monastic buildings and their many smaller stupas continuously reconstructed and renovated, no absolute chronology has been established for the arts of this epoch.

In Laos and Vietnam Theravāda monasteries, with brick stupas, were similarly built and rebuilt of wood. An outstanding stupa is the That Luang at Vientiane, in Laos, founded in 1566 but much restored in the 18th–19th centuries. In Vietnam local variants of Chinese styles were adapted during the Middle Ages to the planning and decoration of palaces and of Confucian, Taoist, and Buddhist temples.

The primitive styles that prevailed in the Philippines were modified by the conversion of various groups—the Moro people, especially—to Islām in the 15th–16th centuries. When, in 1571, the Spanish took control, Manila became the capital of a Spanish colony, and Roman Catholic Spanish art was adopted via Mexico. A local school of church architecture and figure sculpture flourished until the 20th century, when Manila became the centre of a modern commercial society, with its attendant architecture and art.

BURMA

One date is crucial in the art history of Burma: AD 1056. In that year King Anawrahta of Pagan decreed Theravāda Buddhism to be the state religion of all Burma. This signalled the unity of what had been a divided country, consummating tendencies apparent in earlier Burmese history.

6th to 11th century. The only major Burmese art known to scholars is based upon Indian and Ceylonese Buddhist art. In the period preceding Anawrahta's decree there had been three major historical eras in what is now the country of Burma, the first two of which produced Indianized art known to scholars only fragmentarily: the rule of the racially Mon kingdom of the lower Irrawaddy (9th–11th centuries), the contemporaneous dominion of the Pyu people in central and Upper Burma, and the subsequent decisive incursion of racially Burmese people from the northeast (11th century).

Major historical eras

The earliest concrete evidence of Indian culture in Burma is a Buddhist inscription from Pye (Prome) dated *c.* AD 500. This and later inscriptions from the same area were cut probably in the western Mon kingdom, which followed Theravāda Buddhism and was confederated with the Theravāda Buddhist eastern Mon kingdom of Dvaravati (see below *Thailand and Laos*) in southern Thailand and part of Cambodia (AD 6th–12th centuries).

During this same period, in Upper Burma, the people called Pyu, speaking a Tibeto-Burman language and perhaps originating in Central Asia, built cities whose magnificence was known to contemporary compilers of the Chinese Tang dynasty history. In the 8th century one city was recorded as being some 50-odd miles (80 kilometres) in circumference, containing 100 Buddhist monasteries lavishly painted and decorated with gold and silver. The Pyu were in direct contact with northeast India, where various forms of Mahāyāna Buddhism, which embraced philosophies and rituals unacceptable to the Theravāda, flourished; their Ari priesthood was later proscribed by Anawrahta. Their capital city, Śrī Kṣetra (modern Hmawza, near Pye), which was once larger than even Pagan or Mandalay, has been partly excavated. Three huge Buddhist stupas—one 150 feet high—survive there. They illustrate the pattern from which all later Burmese stupas were developed. Enshrining revered relics of Buddhist saints, they consist of tall, solid brick cylinders mounted on shallow, circular, stepped plinths and crowned by what was probably a tapering bell-like pinnacle. Other excavated halls, one on a square plinth with four entrance doors, follow Indian examples. A few Hindu fragments survive as well.

The Pyu were conquered by a neighbouring kingdom, probably before AD 900. During the following century their terrain and cities were infiltrated by the racially Burmese people. These people were of common tribal stock with the Thai and northern Vietnamese and were probably on the move under pressure of the Chinese colonization of their home terrain around the Gulf of Tonkin. They were converted to Buddhism by the Pyu and later by the western Mon; but they never completely abandoned their own original cult of nature spirits, known today as the *nats*. The *nats* are a mixed collection of spirits that act supernaturally, each according to its character. The *nats* were worshipped with orgiastic ceremonies and trance rites of spiritual possession. Certain mountaintops were sacred to them. Even in the 20th century the *nats* exert a powerful influence on the lives of the ordinary people; every village has its own *nat* house—a fragile pavilion built into a tree, after the pattern of the tribal house, and adorned with shreds of coloured cloth, glass, and other offerings. The Buddhist temple in Burma is conceived essentially as an enormous *nat* house, a section of the domain of the

The *nats*

Burmese architecture.
(Left) Library, Pagan, *c.* 1058. (Right) Shwesandaw *cetiya*, Pagan, 11th century.
Louis Frederic—Rapho/Photo Researchers

spiritual located upon earth. And, since the Buddha was adopted as the last and greatest of the *nats*, the same symbols of supernatural splendour that adorn the *nats* adorn the Buddha's images, and a *nat*-like spirituality attaches to the ubiquitous monks in whom the presence of Buddhism is experienced as an everyday reality.

11th century to the present. When King Anawrahta came to the throne, he captured the Mon city Thaton in Lower Burma and carried off its royal family, many skilled craftsmen, and most of the Theravāda monks to his own northern city of Pagan. The king recognized the superior culture of the Mon captives; he established their main form of Buddhism by decree and gave them the task of organizing and civilizing the new united Burmese kingdom and producing for it a Buddhist art. Under Anawrahta's successor, links with the Buddhist homeland were forged. Embassies were sent to Buddh Gaya, in Indian Bihār, and the great Mahābodhi temple there—marking the spot where the Buddha achieved enlightenment—was restored with Burmese money and somewhat in Burmese taste. A smaller copy, with its large rectangular block crowned by the characteristic pyramidal, storied tower, was built at Anawrahta's Pagan. It is here that the greatest achievements of western Mon art—a splendid profusion of architecture and decorative work—are probably to be found. After 1287, when Burma was sacked and garrisoned by the Mongols, new construction at Pagan was virtually abandoned.

Pagan

In Pagan (founded *c.* 849), architecture is the dominant art; except for the big brick icons, mostly ruined, sculpture and painting play a subordinate role. Pagan contains the largest surviving group of buildings in brick and plaster of the many thousands that once stood in various parts of South Asia. The remains at the site, all religious buildings of one kind or another that must once have been surrounded by dense building in perishable materials, are in varying states of preservation. The inscriptions they bear indicate that royal devotees often turned their palaces over to religious use; so it is likely that palace and monastic architecture were very close in style. A few structures still stand that belong to the period before Anawrahta, some of them inspired by Mahāyāna Buddhism and one—the Nat Hlaung Gyaung (*c.* 931)—by Hinduism. Flanking the Sarabha Gate is a pair of small *nat* shrines with pointed, open windows—the earliest in Burma, perhaps 9th century.

The library, built about 1058 to house the books of one of the Buddhist monasteries, is one of the most important buildings in Pagan. It is rectangular, with a series of five stepped-in, sloping stone roofs crowned by a rectangular tower finial. The concave contours of the roofs are characteristic of much Burmese architecture. The eaves and corners of all the tiers are adorned with the typical Pagan flame ornament, or antefix.

The *cetiya*

There are other buildings of the same general type among the ruins of Pagan. Far the most numerous and important, however, are the buildings—called *cetiyas*—that combine the attributes of stupa and shrine. These have a history and a line of evolution of their own, which can be traced from the Pyu stupa to the huge structural temple. The normal stupa, derived from the early medieval Indian form, is a tall structure consisting of a solid dome set on a tiered square plinth (often with miniature stupas at the corners) around which the faithful may perambulate. The dome is surmounted by a *harmikā*, which resembles the small railed enclosure found on the oldest Indian stupas. In Burmese stupas, however, the *harmikā* becomes a decorated cubical die, above which is a circular pointed spire; in memory of its distant origin in India, the spire is horizontally flanged (rimmed) with moldings in a series of honorific umbrellas of decreasing size. In later practice *harmikā* and umbrella spire become a single architectural unit. The Burmese stupa dome, based on the tall, cylindrical Pyu prototype, has a spreading concave foot resembling a bell rim. The Lokananda and Shwesandaw at Pagan are two well-known examples. Because in recent times they have been coated with plaster, the finely detailed brick carving characteristic of early Pagan architecture has been obscured. Such carving is beautifully exemplified in the Seinnyet temple at Myinpagan (11th century).

Anawrahta's type of *cetiya* followed the general form of the early Pyu stupa. The main point of evolution was in the progressive elaboration of the terraced plinths on which the dome stands. The plinths became virtually sacred mountains, with a series of staircases running from terrace to terrace up each of the four sides. Perhaps inspired by vanished work in contemporary late-11th-century India, the Burmese began to open up the interior of the terraced base of the stupas with wide corridors and porticos, converting it into a roofed temple. The cylinder of the stupa dome was carried down through this temple space to its floor. Four large Buddha icons were added to the lower part of the dome, facing the four directions. Once this conception had evolved, it was possible to create around the central stupa a broad circuit of roofed enclosures, which from the outside would still suggest the traditional pattern of the stupa standing on its raised terraces, while the interior could be used for ceremonial, as in a true temple. Sculpture and painting, decorating the internal halls, corridors, and doorways, recounted the life of the Buddha and presented the example of his previous virtuous incarnations. The most famous example of this type of *cetiya* is the great Ananda temple at Pagan (dedicated 1090). It is still in use, unlike most of the old temples there, and so is kept in repair; it is painted a blazing white with lime stucco—which has, of course, obscured the finer detail of its old architecture. Its plan is square, with a broad, four-pillared porch hall added to all of the four doors in the four faces of the square. Its tower is a curvilinear pyramid resembling eastern Indian Hindu temple towers, and its enormous brick mass is pierced with two circuits of vaulted corridors. The sloping, curved

terrace roofs have an elegant overall concave profile and flame antefixes along all the eaves.

As time went on, Burmese brick and stucco architecture developed principally through the stiffening of masses into rectangular blocks and through the elaboration and often the coarsening of its ornament. The 13th-century Gawdawpalin temple at Pagan, for example, consists of a rectangular hall with a large closed entrance porch; the hall is surmounted by a tall but narrow second story whose decoration repeats that of the lower story; the whole building is crowned by a four-faced tower with a curved profile. Multiple moldings and decorative motifs are used as outlining elements and the doors are framed in elaborate upward-flaring hooded porches.

Until the Mongol conquest in 1287 much excellent work seems to have been done at Pagan. It is, however, impossible to form an adequate idea of the older styles of temple architecture at other sites in Burma, such as Rangoon or Mandalay. Whereas most of the temples of Pagan were abandoned early on, so that even though they may be ruined they show their original characteristics, temples in modern cities have been repeatedly and drastically restored. Old stupas may have as many as eight successive casings of brick and stucco; temple walls and doors are constantly torn down and rebuilt; and stucco surfaces may be renewed almost annually. Such attentions to a religious building are popularly regarded as acts of merit; thus, revered architectural monuments suffer continually from well-intentioned but disastrous renovation. At the big stupa sites huge numbers of pagodas are constantly falling into decay and new ones are being built at great speed. Among them are variants, whose evolution cannot at present be traced, on the basic pattern of the long tapering bell, with a variety of transverse moldings, standing perhaps on a recessed plinth. Many are covered quickly with extravagant and gross stucco ornament. Ornate flaring porches and flame finials are added to gates, wall ends, and eaves corners. A tapering slenderness is the outstanding characteristic of all the different types.

Stupa sites in Rangoon and Mandalay

The monastic architecture—patterned on the hall, with its elaborate doors—that surrounds the great stupa sites of Rangoon and Mandalay is mainly in wood, built by simple pillar and architrave construction. The roofs are steeply gabled, with multiple gables riding over each other on immense carved pillars in the larger halls. The angles between pillar and architrave and the edges of roof gables, tiers, and terraces are filled with flamboyant cartouches (scroll-shaped ornaments) of pierced work, often lacquered and gilt; thus, the whole building may be smothered in repetitive ornamental curlicues. All this ornament has an otherworldly or spiritual significance. Other stupa sites in Burma, where less money has been spent and less ornament added to the buildings, may be more beautiful to the modern eye, with only a few flamboyant antefixes pointing the gables and punctuating the eaves. All over Burma similar buildings can be found; but, while many have been listed, they have been scantily surveyed, and no real study of their complex history has yet been attempted. There may well be a substantial Chinese influence in the construction of some of the wooden halls and pavilions.

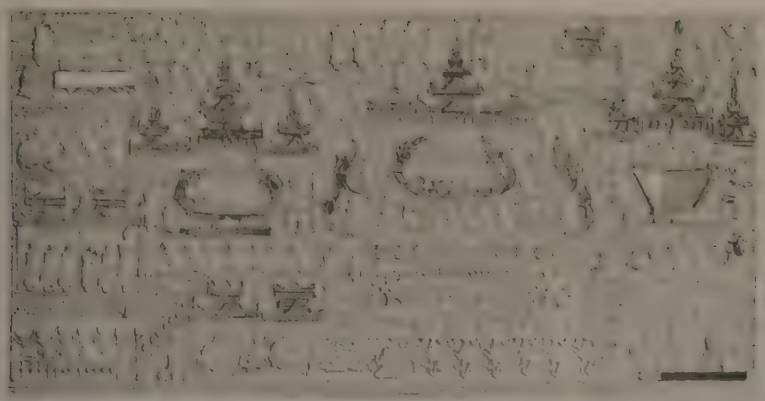
Paintings and sculpture in Theravāda Burma do not seem to have reached the same heights of achievement as in other countries of Southeast Asia. They do not show the same originality and sense of life. The temples of Pagan contain the best examples, although even these are highly schematic, reminiscent in design of eastern Indian Buddhist manuscript styles. A number of early buildings at Pagan contain fragmentary terra-cotta (fired clay) reliefs or scraps of wall painting whose individual figures display some of the sensuous charm of their Indian prototypes (it is quite likely that Indian artists worked there). The 12th-century terra-cotta panels from the elaborate facings of the Ananda temple, however, show the beginnings of the petrifaction that overtook later Burmese figurative art. Both in reliefs and in wall paintings, the figure compositions are reduced to schematic groups of the minimum number of standard human and celestial types needed to tell a moral story, without any infrastructure of significant form and execution. The colossal Buddha images enshrined in the

(Top) Ananda temple, Pagan, Burma, dedicated 1090. (Bottom) Section of the Ananda temple.

(Top) Louis Frederic—Rapho/Photo Reseachers, (bottom) from P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

temples were usually built of brick and finished in stucco, gilded and ornamented. Such work is still done at high speed today. The technique is not a flexible one, and the emphasis of Theravāda Buddhism on the exact imitation of ancient Buddha images gave the Burmese no aesthetic incentive to develop the expression of their figures or compositions. The repeated heavy gilding and repainting of older icons has almost entirely destroyed any formal vitality they may once have had.

From about 1700 to 1850 Burma excelled in the decorative arts, whose forms continually recall those of Theravāda Ceylon. Burmese woven silks and embroideries are well-known. The carved wooden screens, panels, and brackets used inside temple halls, many devoted to representing the *nat*s and the population of the spirit world, have benefitted from being outside the strict canon of Theravāda Buddhist orthodoxy. The figure types follow fluid, slightly "boneless" conventions derived from classical Indianizing dance postures. The decorative goldwares and silverwares, which use much stereotyped decorative scrollwork, are also based on standard Indianizing iconography. Perhaps the most aesthetically satisfying works of the Burmese sculptors are the reliefs ornamenting the *sū-*



Front view of a Burmese *sūtra* chest, wood and gesso, probably 18th century. In the Gulbenkian Museum of Oriental Art, the University of Durham, England. 63.5 × 120.6 cm.
By courtesy of the Gulbenkian Museum of Oriental Art, the University of Durham, England

tra chests that were used in monasteries to store the sacred texts of Buddhism. The gilt gesso (paste used for making reliefs) facings of these chests carry the schematic style of relief sculpture beyond its normal aesthetic limits. This is accomplished by the way they are compelled to set off their figures as an intelligible scheme of thin raised lines and inlay against a plain ground. The forms of the fine lacquer bowls and boxes used in monasteries, decorated only superficially with painted ornament, show the underlying formal sense of the Burmese at its clearest.

Regional Burmese art

Interesting regional types of Burmese art are those of the Shan and Karen peoples, who live in the relatively remote northern hills. These areas have often produced extremely beautiful types of domestic and religious architecture, made of wood, on stone bases. They are a simpler and more austere version of the ancient pattern that underlies the halls and pavilions of the more sophisticated recent southern temple buildings, with their steep, gabled roofs. The peoples of the north also produce a variety of decorative arts. Notable among them are the textiles, which are characterized by banding, checkering, and triangular counterchanging of brilliant colours set off against black. The woven shoulder bags, particularly, are well-known in the West.

THAILAND AND LAOS

Dvaravati Mon kingdom: 6th to 11th century. Archaeology has recovered in central Thailand substantial glimpses of the magnificent early layer of Indianized culture, which includes a religious art that was produced between the 6th and 11th centuries by the eastern Mon kingdom of Dvaravati. The art was created predominantly to serve Theravāda Buddhism. Remains of Dvaravati architecture so far excavated include stupa bases: notable examples include the Wat Phra Meru in Nagara Pathama (Nakhon Pathom) and others at Ku Bua and U Thong, some of which have elephants supporting their bases, following a pattern that originated in Ceylon. The plinths of Buddhist assembly halls, which existed near the solid monumental structures, have also been discovered. Many terra-cotta and stucco fragments of decorative surface designs and celestial figures have also been found. The Wat Pra Meru, on a plan similar to that of the Ananda temple at Pagan in Burma (see above *Burma*), probably antedates the latter's foundation (*c.* 1090). It is likely that many other ancient monuments are encased in later stupas that are still being used for religious purposes, for it was probably customary not to destroy an old sacred monument but to encase it in a new shell, maybe several times over, and perhaps to construct a small external replica of the encased original alongside.

At many sites, especially Lop Buri, Ayutthaya, and U Thong, fine Dvaravati sculptures have been found among the architectural remains. Particularly important are the seated and standing Buddha figures in stone and bronze. Many of the faces have characteristic Mon features, with lips turned outward (everted) and downward-curved eyelids marked by double channels. Some of these Dvaravati images may well have furnished models for later Khmer art in Cambodia.

Dvaravati sculpture

Dvaravati sculpture shows close relations with several Indian styles, notably those of Amarāvatī, Gupta, post-Gupta, and Pāla Bihār. It also was probably influenced strongly by the art of the enigmatic kingdom of Śrivijaya in Sumatra, as well as by central Javanese types (see below *Indonesia*). One outstanding masterpiece from Chaiya, of Dvaravati date, may well be a work produced in Śrivijaya. It is a bronze torso and head of a *bodhisattva*, for which a mid-8th-century date is suggested. The body and face are modelled with a plastic and delicate sensuousness; and the elaborate necklaces, crowns, earrings, and armlets are beautifully chased (decoratively indented by hammering). The Śrivijaya origin is made more likely by stylistic reminiscences of the sculpture of contemporary Indonesia, which was also under Sumatran inspiration.

Khmer conquest and Tai immigration: 11th to 13th century. In the 11th century Dvaravati was captured by the Khmer of Cambodia and became a province of their empire. A number of Khmer shrines, probably intended as focuses of the Khmer Hindu dynastic cult, were built in Siam (Thailand). At Phimai (Bimaya) was the most important full-fledged Khmer temple, where one of the personal cult statues of the Khmer king Jayavarman II (see below *Cambodia and Vietnam*) has been found, together with bronze images, some of Tantric Buddhist deities. At Lop Buri the Phra Prang Sam Yot is perhaps the best surviving example in brick and stucco of Khmer provincial art in Thailand, its tall towers having complex rebated (blunted) corners and its porticoes high, flamboyant pediments (the triangular space used as decoration over porticoes, doors, and windows). Wat Kukut, at Lamphun, built by a Dvaravati Mon king *c.* 1130, represents an adaptation of the Khmer stepped-pyramid temple base as pattern for the temple itself. The niches on its terraces are filled with images in a deliberately archaistic revival of the old Mon style.

Domination by the Tai peoples

During the period when the Khmer were taking over the southern Mon region of Thailand, the northern region was falling under the domination of immigrant racially Tai peoples. The Tais were a branch of the migrating population who invaded Burma as the Burmese and of the Sinicized Vietnamese who were then pushing southward into what is now Vietnam. The Tais seem to have professed an animist nature religion, resembling the early form of the Burmese cult of the *nats* (see above *Burma*). This whole group of peoples originated most probably as a tribal population in the region of Tongking and Canton. In the course of their southward migrations they probably played an important role, as yet unclear, in a kingdom called Nanchao, in what is now the Chinese province of Yunnan. The rulers of this kingdom seem to have followed a Mahāyāna form of Buddhism, including the cult of a *bodhisattva* as personal patron of the king. Several smallish bronze icons of a *bodhisattva* with a nude torso and a strap round the upper belly are known from Nanchao, in a style reminiscent of the later Pallava art of the

east coast of peninsular India. The date of these images is still uncertain. Tai kingdoms were gradually established further and further south. Some of their tribes gained experience of administrative techniques by living within the boundaries of the Khmer Empire, with their own chieftains under Khmer officials. When the Khmer power was broken in the 13th century, the Tai moved into central and southern Siam, intermarrying with the Mon.

The Tai people normally built in perishable materials, wood and bamboo in particular. Their animist religion, which has no canonical group resembling the Burmese *nats*, is still very much alive today. The spirits of trees need to be pacified, and the ancestors can be powerful helpers. Shamans, in a state of trance, make contact with the spirit world to perform good or evil magic. In the wooden high-gabled houses of the northern Tai (Chiengmai province), even today, ornate lintels are carved with floral relief designs to sanctify and potentiate the inner domestic part of the house where the domestic spirits live. The animist religion gave ground partially to Buddhism, which was gradually assimilated among the people, and at some date, as yet uncertain, was adopted by the greater Tai kings as a dynastic religion. With the spread of Buddhism a special religious architecture in brick and stucco was established.

Division of Siam

The Thai kingdom: 13th to 17th century. During most of its history, Thailand has been divided into two fairly distinct regions, a northern and a southern, the capital of the north at Chiang Mai, the capital of the south at Ayutthaya. Between the two lies the great trade-route city of Sukhothai, possession of which fluctuated between the north and the south. Sukhothai seems to have been the principal focus and source of Buddhist culture in Siam, for it retained direct touch with Ceylon, which, after the decline of Buddhism in India in the 12th century, became the principal home of Theravāda Buddhism. By the 15th century the difficult art of casting large-scale Buddha figures in bronze had been mastered in the north of Siam, as well as in the south.

Sculpture. The Thai kings made repeated attempts to "purify" their conservative Theravāda strain of Buddhism, importing patterns of art along with texts and learned monks from Ceylon and trying to wean their people from worship of the spirits. To retain the greatest spiritual potency, Buddha icons in Thai temples had to be as close in type as possible to a great original prototype that Buddhist tradition erroneously believed had been made during the lifetime of the Buddha; in practice, this meant the types the local craftsmen knew as the oldest and most authentic. There were at least three major successive efforts by Thai kings to establish and distribute an "authentic" canon for the Buddha icons, which were their prime artistic concern. Each type that became canonical and was known to be magically effective was imitated repeatedly. For it was regarded as an act of merit simply to multiply images of the Buddha, whether they were to be installed in temples or not; hence, in addition to icons, enormous numbers of small images—made of many materials, from bronze, silver, stone, and wood to terra-cotta—were kept in temple storehouses. The images followed canonical patterns established for the major temple icons.

Since their work had to be as similar as possible to the oldest sacred images of which they knew, the Buddhist sculptors in Siam adhered to strict formulas and diagrams; artistic development was never a part of their purpose, though of course gradual change did occur. There is no tradition in Theravāda Siam in any way resembling the traditions of Mahāyāna art in, say, Cambodia or Indonesia, which encouraged artists to explore the possibilities of their mediums to express developing religious conceptions. Thus, Thai Buddhist sculpture consisted almost entirely of careful repetitions of the standardized types, which tended naturally, despite the artist's desire to capture an authentic sense of style, to lose their older vitality. It also happened that the three main canonical patterns often lost their individuality, blending into each other. Perhaps the best works were made in the 15th century, but work of high quality was still being done in the 16th and early 17th centuries.

Canonical types of Buddha image

The first canonical types were the Sukhothai, which seem to have been evolved in the trade-route city of Sukhothai as an attempt to capture the quality of early-medieval Ceylonese images and elements from Dvaravati sculpture. The developed versions of these types are marked by an extremely smooth, rounded modelling of the body and face, without any clearly defined planes. The outlines of hair, eyebrows, lips, and fingers are elegantly recurved, or S-curved, and the head is crowned by a tall, pointed flame finial. The entire figure gives an impression of great elegance. Full-fledged Sukhothai images of the full-round walking Buddha—an original Sukhothai invention—emphasize a kind of swaying, sinuous, boneless grace in the execution of the legs and arms. One of the most impressive colossal images of the type is the brick and stucco icon at the Wat Mahathat, Sawankhalok, another Sukhothai technical forte, dating probably to the 14th century. This type of image remained the most popular in Siam; an enormous number of imitations, of all dates, are preserved, many in Western collections.

Perhaps the Buddha types most successful aesthetically were those called after U Thong. They were produced originally in the southern capital of Ayutthaya, which took over Sukhothai in 1349, and represent a fusion of the Sukhothai types with vestiges of Khmer and Theravāda Dvaravati traditions, whose Buddha types had been marked by a strong Mon sense of squared-off design and cubic volume. The latter may have been influential because they seemed to incorporate an older and more authentic tradition, since they were based upon patterns developed in eastern India, the true homeland of Buddhism. In the U Thong style the sinuous linear curves, loops, and dry ridges of the pure Sukhothai patterns are suppressed, and genuine modelling, with clearly defined

By courtesy of the Breezewood Foundation, Monkton, Maryland

"Buddha Calling the Earth to Witness," Thai, bronze, Sukhothai high classical style, 14th century. In the collection of Prince Chalermbol Yugala, Bangkok. Height 94 cm.

planes and volumes, appears. In the northern kingdom a crude version of the Sukhothai type gained currency in the late 14th century. When, in the middle of the 15th century, King Tiloka of the northern kingdom reestablished contact with Ceylon, images seem to have been imported directly from that country. They must have shown clearly how far the Sukhothai types had departed from the type used in the Buddhist homeland, because the third Siamese icon pattern, known as the lion type, attempted to recapture the stern simplicity of the genuine Sinhalese images. Most of the best examples were made between 1470 and 1565. Limbs and bodies are given a massive cylindrical strength, and the Sukhothai elegance

is eliminated. It seems, however, that the native Thai genius is for the sinuous and unplastic curve, which may have expressed for them the same spiritual unworldliness as it did in Burmese ornament. Thus, in later examples reminiscent of the lion type, the curvilinear patterns of the Sukhothai style reassert themselves with more or less emphasis; and by the end of the 16th century the lion type had lost its distinguishing features and merged into the run of Sukhothai patterns.

Architecture and painting. There are as yet few results of authenticated research available concerning the history of architecture during the early period of Thai supremacy. Many monasteries contain stupas, or *cheddis*, that either originated or were renewed in this period; but most of the monasteries themselves have been repeatedly overworked. Building complexes seem to have developed by accretion, rather than by the studied working out of space articulations. The oldest building in Ayutthaya, dating from the early 13th century, is the Wat Bhuddai Svarya, a towered shrine, approached by a columned hall. From the late 14th century onward, Sukhothai influence seems to have predominated everywhere. The architectural types included a bell-shaped reliquary stupa with a circular flanged base and onion finials, reminiscent of combined Ceylonese and Burmese patterns; a stupa raised upon a cylindrical shrine as its drum; and a shrine with a plinth faced with images (usually later additions) above which rise one or more pyramidal towers reminiscent of the tower of the Mahābodhi temple at Buddh Gaya in Bihār, India. An example of the third architectural type is King Tiloka's late-15th-century Wat Chet Yot at Chiang Mai, which has one large and four smaller pyramids mounted on a main block.

14th-century architectural types

Louis Frederic—Rapho/Photo Researchers

Wat Chet Yot, Chiang Mai, Thailand, late 15th century.

The Thai kings also adopted something of the personal funeral cult of Khmer Angkor (see below *Cambodia and Vietnam*), for a custom grew of building bell-shaped brick stupas—which had earlier been used only for the relics of Buddhist saints—as the kings' tombs, each approached by a colonnaded hall and surrounded by smaller stupas or shrines. In many of the brick and plaster or wooden monastic buildings of more recent centuries, such as the Wat Po in Bangkok, one can trace the distant influence of the Khmer styles of Angkor. Tall, gabled roofs, with steps and overlaps, the gables adorned with flame finials, are typical, exemplified by the Water Pavilion at Bang Pa-in.

Water Pavilion, Bang Pa-in, Thailand, 1294.

Luc Bouchage—Rapho/Photo Researchers

Thai painting of the early period (13th–16th centuries) demands a great deal more research and study than it has yet received. Although it is, of course, devoted to the canonical iconography of the Theravāda, its fluent and relatively unschematic outline shows that it retained much of the original inspiration visible in the earlier work at Burmese Pagan (see above *Burma*). The oldest examples of Thai painting are the much-ruined frescoes in the Silpa cave, Yala, and some engraved panels from Wat Si Chum, Sukhothai, dated to 1287. Later paintings (dating to the 1420s) in the inner chambers of the Wat Rat Burana and Wat Mahathat at Ayutthaya show strong Chinese and perhaps Khmer influence in their high perspectives and landscape backgrounds with animals, combined with the native Thai clear outlines and bright, flat colours. By the 17th century at, for example, the Wat Yai Suwannaram at Phet Buri, large mural compositions—such as an elaborate scene of demons worshipping the Buddha—were being undertaken. In this later painting, theatrical stereotypes from the Thai dance-drama exerted a strong influence in the rendering of figures.

18th century to the present. In the 18th century the Burmese invaded and conquered Siam. The Burmese king—in expiation, it is said, of his war guilt—ordered the construction of many Buddhist buildings in the current Burmese style (see above *Burma*). These made their impact on Thai art, and the gaudy gilding and inlay characteristic of late Burmese ornament were widely adopted. When the capital was moved to the present Bangkok, in 1782, no substantial artistic development took place, though large pagodas were built and filled with rows of images, many in gilt wood. A highly ornate interpretation of older, airily flamboyant Burmese decorative styles, featuring curved "oxhorn" projections, blunted the edge of architectural and sculptural quality. Without exception, the new large-scale icons were dull and inferior works of art; and the monstrous guardian figures of spiritual beings and lions decorating the major shrines are fantastic rather than aesthetically valuable. In the painting of wooden panels, some of them votive, and of historical manuscripts, the Thai retained a good deal of their older vigour. The figures illustrating legend and history are based upon the unworldly stereotypes of the court dance.

Burmese influence on Thai art

In addition to the incorporation of European motives, many buildings and their ornamentation in Bangkok have a strongly Chinese flavour. This is attributable partly to the influence of the large expatriate Chinese population

Sawankhalok, Sukhothai, and Chiang Mai pottery

living there and partly to the influence of earlier expatriate Chinese craftsmen. The early 20th-century Pathamacetiya at Nagara Pathama (Nakhon Pathom), which is entirely orange, is a fine example of the many *cheddis*. Some tiles were certainly imported from China, but others were descendants of the fine pottery (of basically Chinese inspiration) that was produced at the kilns of Sawankhalok during the 14th and 15th centuries by expatriate Chinese craftsmen. This pottery imitated in its own materials Chinese Yüan dynasty (1279–1368) Tzu-chou and celadon wares (stonewares and porcelain with a glaze developed by the Chinese) with underglaze ornament and blue or brown painted decoration. Similar wares were made in the 15th century at kilns at Sukhothai and at Chiang Mai. Some of these pieces are, in their own idiom, as fine as native Chinese work. Later, during the 18th and 19th centuries, somewhat garish, flamboyant Ayutthaya figure designs in polychrome were applied to rice bowls and other vessels.

Laos. The kingdom of Lan Xang (Laos) was founded in the mid-14th century and ruled by Buddhist Thai. At the northern capital, Luang Prabang, the influence of the northern Thai city of Chiang Mai predominated; in the southern capital, Vientiane, a mixture of Ayutthaya and Khmer motives prevailed. In Laos there is no stone and little brick architecture. The most impressive single monument, the brick and stucco That Luang in Vientiane, founded in 1586 but much restored, is a stupa, shaped as a tall four-faced dome on a square plinth enclosed in a court; the dome is crowned with an ornate spire and encircled by a row of similarly shaped spires. The architecture of monastic halls also follows the Thai pattern; very steep multiple-gabled roofs, gently curved and overhung with long eaves, are carried on brick or wooden pillars and adorned with flame finials. Buddha figures, preserved in some of the monasteries, are based on northern Thai versions of Sukhothai types; some may be as early as the 17th century. The schematic paintings on monastery walls are in versions of the later Thai styles. In the northwest a strong influence from late Burmese art can be found in Buddhist images made to serve a religion that was far closer to the original Thai animism than to true Buddhism.

CAMBODIA AND VIETNAM

Paleolithic tools similar to types found in India have been found in Cambodia (Kampuchea) and Vietnam; and it is possible to trace the movement of population or culture groups, some of whom probably migrated onward by sea from Southeast Asia into the islands. The important group of speakers of Mon–Khmer languages may conceivably have been the people who produced the megalithic monuments in Cambodia and Laos, which include colossal stone burial urns, dolmens, and menhirs, perhaps associated with the many circular earth platforms as yet unexcavated (see above *General development of Southeast Asian art*). Probably contemporaneous, at least in part, with the Neolithic Mon–Khmer culture is the culture known by the name of its richest, most northerly site, Dong Son, on the coast of the Gulf of Tonkin in northern Vietnam. It seems probable that the chief influences on this culture came from southern China. Many sites, ranging in date from about the 4th to the 1st century BC, stretch southward from the coast of Vietnam, as far as northern New Guinea. The islands of Indonesia and parts of Malaya may have been the principal location of the Dong Son culture.

Bronze drums

The most impressive bronze objects produced by this culture are large drums, which seem sometimes to have been buried with the dead. Splendid examples have been found in Java and Bali (see below *Indonesia*). These and many other bronze objects, such as superb funeral urns with relief ornament based on squared hooks, lamp holders, dagger hilts in the form of human figures, and other weapons, are of extremely high quality. Their ornament was produced by the Chinese casting technique of incising the patterns into the negative mold that was to receive the molten bronze; much of it suggests a parallel version of contemporary Chinese ornament of the Ch'in period (221–206 BC). From the figures and objects represented in this bronze work, it seems that the Dong Son culture had much in common with that of some of the peoples

Lamp holder from Lach Truong, Vietnam, bronze, Dong Son culture. In the National Museum, Hanoi. Height 33 cm.

By courtesy of the Fine Arts Conservation, Ministry of Culture, Hanoi

of the Melanesian islands today. The culture knew large seagoing canoes, houses similar in structure to those still common among peoples of Melanesia, and ceremonies that the Melanesians might recognize. It is probable that one group of their descendants, which retained its identity, is known to the history of this region as the Cham (see below *Vietnam kingdom of Champa*).

Although many peoples isolated in the densely forested uplands also retained a tribal identity, by far the most important art was produced in the two Indianizing empires: Khmer, in Cambodia, with its linear predecessors the kingdoms of Funan and of Chenla (names they were given by Chinese historians), and the Cham, in Vietnam.

Cambodian kingdoms of Funan and Chenla: 1st to 9th century. Funan, which was in existence by the 1st century AD, was the earliest of the kingdoms that arose along the lower reaches of the Mekong River in response to Indian ideas. Its influence probably extended over long stretches of the coast of the Gulf of Siam, even as far as southern Burma, and corresponded with the range of the Mon peoples. Lying on the natural focus of land and sea routes linking eastern India and southern China to the islands of the South Seas, its geographical situation was ideal for a kingdom whose wealth was based on trade. At Funan sites even Roman, Ptolemaic Egyptian, and Sassanian Persian objects have been found, giving an idea of the extent of its trading interests.

The founder was probably a Brahmin trader from western India; for a local legend describes how the first king, a Brahmin, married the daughter of a local serpent deity, so establishing the ruling family. Serpents (*nāgas*) in Indian mythology are the spiritual patrons of water; and the basis this kingdom laid for later kingdoms in the same area was an elaborate system of waterworks, canals, and irrigation channels controlling and distributing the waters of the Mekong River. Contemporary Chinese accounts refer to cities with splendid wooden buildings, carved, painted, and gilded. But nothing remains, save a few foundation piles. Probably during the 6th century AD the kingdom called Chenla was established in the upper-middle reaches of the Mekong River, in what is now Laos. The kings who ruled in Chenla were descended from the kings of Funan and took over much of the Funan domain. It seems that disastrous floods had finally ruined Funan, which had previously suffered from Indonesian aggression, and that the shift of power to Chenla represented a recognition of temporarily insuperable geographical difficulties.

Stone sculptures

Culturally, Funan and Chenla are continuous. Their artists produced some of the world's greatest stone sculptures, most of which are large, freestanding icons, carved in

sandstone. Intended to be installed in brick-built shrines, none of which survive, they usually represent the two major deities of Hinduism, Śiva and Vishnu. Sometimes both deities are combined into a single figure called Harihara; the right half of the body is characterized as Śiva, the left as Vishnu. A few examples of other figures are known, including some magnificent images of goddesses. The style of these sculptures is marked by an extremely smooth, continuously undulating surface, given strength by a system of clear, broad frontal planes and side recessions related to the foursquare block. Such images were meant to demonstrate the power and charm of a heavenly prototype to whom an earthly king appealed for his authority. The earliest images belong to the 6th century, and the series continues into the 9th.

In later Khmer times each king and sometimes each member of a royal house had statues of himself or herself in the guise of a patron deity set up in the family temple precinct. That the same custom prevailed in 6th-century India, particularly in the southeast, suggests that some of the early Funan and Chenla sculptures may have served the same function. A number of figures are Indian in style—some more markedly than others, which is probably more than a matter of date; for it is quite likely that Indian craftsmen occasionally travelled into this region to work. The style of the greatest of these early sculptures, however, is not Indian at all.

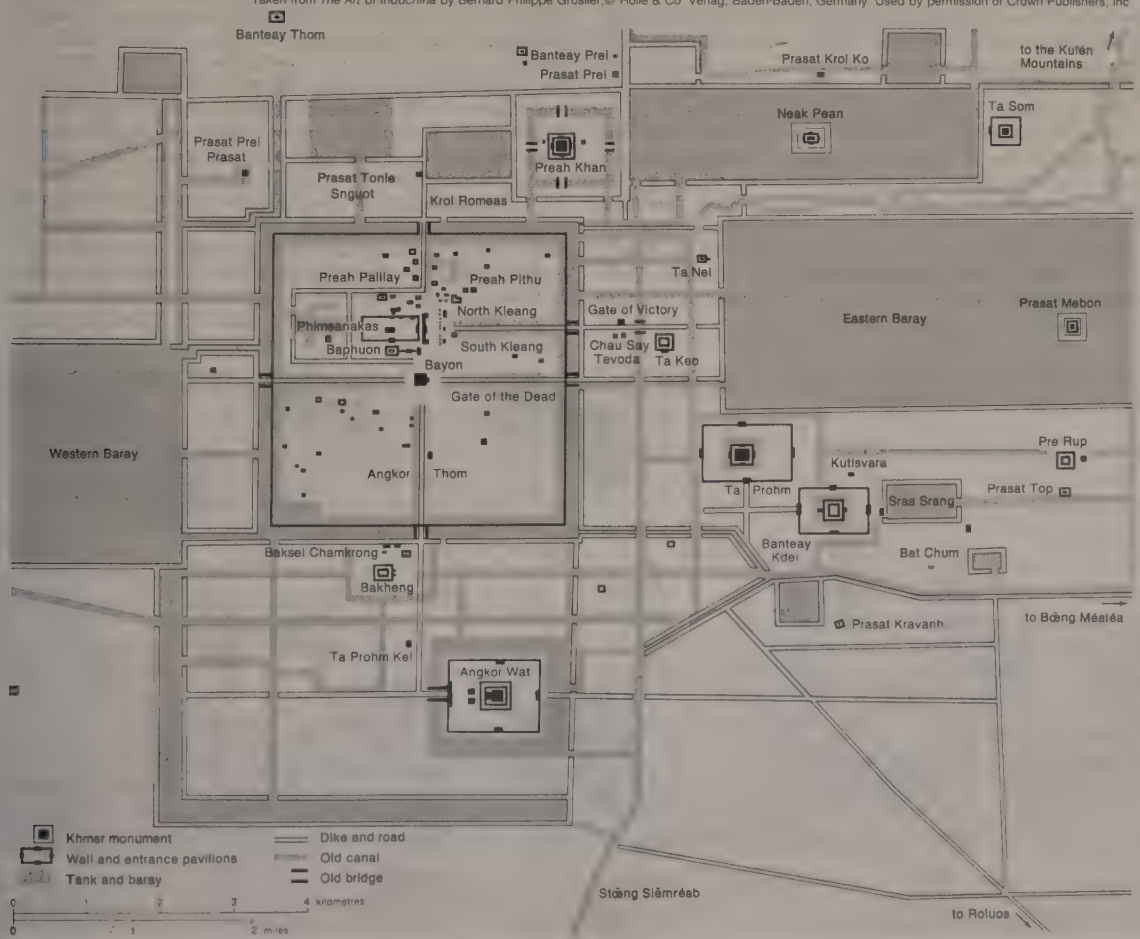
Sandstone lintels

Similarly non-Indian are the magnificent sandstone lintels made for the doorways of the vanished brick shrines. Although distantly related to Indian prototypes of the 1st and 2nd centuries AD, they appear as full-fledged Indochinese inventions and may well have been developed in combination with a native conception of the lintel as a special attribute of the spirit shrine (see above *Thailand and Laos*). They are carved in relief with designs based on a pair of monsters, one at each end, which are linked by an ornate arched or lobed beam. The beam is adorned with figures inside foliate plaques, a long sequence of elaborately carved swags of jewels hanging beneath them.

Among the Funan–Chenla sculptures are a few Buddhist icons executed in sandstone, markedly less sensuous than the Hindu figures and close to the styles of Dvaravati (see above *Thailand and Laos*), though a number of small Buddhist bronzes representing *bodhisattva*s approach the delicacy of the Hindu work.

Kingdom of Khmer: 9th to 13th century. Late in the 8th century the kingdom of Chenla declined politically, perhaps because of dynastic disputes with the rising power of Indonesian kings, who were themselves also descended from the original royal dynasty of Funan. It seems that the Indonesians gave some assistance in establishing a new kingdom in the northern part of what had been the territory of Funan. In 802 a Khmer king, who took the title of Jayavarman II, established his capital near Phnom Kulen, about 20 miles (30 kilometres) from Angkor. It was a rather unsuitable place for an administrative capital, but it was a mountain, and the peoples of Southeast Asia have always believed that gods and spirits dwell on mountaintops. The image of the sacred mountain thereafter remained the inspiration for all the later architecture of the Khmer around Angkor. Jayavarman, who built other temples in the vicinity, seems to have revived the Chenla style. A distinctively Khmer art, however, began to emerge under Indravarman I (877–889), who expanded the boundaries of the Khmer kingdom and finally settled its administration. Most important of all, he developed the initial plan of the colossal city of Angkor, whose mysterious ruins, lost in dense jungle until very recently, have tantalized Western travellers for three centuries.

Angkor was not only a city; more important, it was an immense technological achievement, from which the agricultural prosperity of the whole Cambodian plain derived. This plain was well watered naturally, but its rivers were subject to strong seasonal fluctuations. Controlled,



Plan of Angkor.

Technological significance of Angkor

they were capable of producing an enormous increase in fertility. Angkor was thus essentially an elaborate system of artificial lakes, canals, and radiating irrigation channels that watered a huge acreage of rice paddy; and it was the basis for the strength and prosperity of the Khmer Empire. Since Angkor itself was the technical source of the life-giving agricultural water controlled by the king, it was regarded by the Khmer with religious reverence. Its temples and palaces were an expression of that reverence and at the same time an essential part of its supernatural mechanism. Royal intercession by numerous ceremonies, some of which re-enacted the primal marriage of Hindu divinity and native earth spirit on the pattern of ancient folk cult, ensured the continuing gift of the waters of heaven. The king, an earthly image of his god, was the intermediary who ensured that his kingdom would continue to receive divine benevolence in the form of water in controlled quantities. Courtiers played roles at once religious and administrative for the king, who believed that after his death he would be united with his patron deity. Dedicatory statues were often set up in his chief temple to commemorate his divinization.

In order to conform with mountain mythology, the Khmer kings built themselves a series of artificial mountains on the Cambodian plain at Angkor, each crowned by shrines containing images of gods and of themselves, their family, and their ancestors. The huge platforms of earth on which these buildings were founded probably consist of the soil excavated in forming the lakes, moats, and channels that not only divided up the city but also provided an easy means of transport. The temple mountains, like the city itself, are oriented east to west, the main gates facing east. Each king strove to outdo his predecessor in the height, size, and splendour of his temple mountain. The earlier ones, therefore, are relatively small, though beautiful, while the later ones, such as Angkor Wat and the Bayon, are of stupendous size.

The basic pattern of the Khmer temple mountain

In the basic pattern of the Khmer temple mountain the principal overall enclosure, which is square or rectangular, is at ground level. Within it the artificial mountain rises through a series of terraces and at least one further enclosure wall toward a flat summit. On the summit stands either a single shrine or a group of shrines, often a quincunx—five shrines, one at each corner and one in the middle of a square. Arranged along the terraces or within the enclosures there may be further shrines, whose arched doorway pediments refer to the rainbow bridge between heaven and earth. There may be other long buildings, perhaps used as libraries or administrative offices. A principal staircase runs directly up from the east gate to the summit, and sometimes subsidiary staircases run up from other gates at the cardinal directions.

The architecture of the shrines themselves is relatively simple; it is based upon patterns invented in India, though the ornament of the shrines is often highly developed and characteristically Cambodian. Fundamentally, each shrine consists of a cell whose internal space is cubic and whose external walls are marked by moldings at top and bottom. The shrine is roofed by a pyramidal tower composed of a series of similar but diminishing tiers, each of them a compressed version of the exterior pattern of the main shrine volume. Depending on which Indian pattern is followed, the cell has one main door with an elaborately carved portal or, if the plan is cruciform, four entrances. The earlier shrines were built of brick, most commonly with stucco ornament and figures on the outside. The later shrines were built of stone, with all their ornament and figurative sculpture carved in relief. The moldings on the roofs of the shrines and the decoration of the roofs of many of the subsidiary buildings are extremely elaborate. There are long panels of dense foliate ornament, and the niches in which the sculptured relief figures of celestials are set are framed in flamboyant ogival (contoured like a pointed arch) moldings crowned by no less flamboyant foliate ornament; the smaller architectural features, such as niche pilasters, are elaborately carved and molded. The figures themselves wear gorgeous jewelry and chignons. The massive stone icons that survive in some of the shrines and on the terraces do not have the subtlety or strength of the Funan–Chenla sculptures. Instead, they have an inflated massiveness, intended, no doubt, to make them awe-inspiring. Among the lesser relief figures of celestials, which decorate the walls of the shrines, one finds a more sensuous touch; for many of these celestials represent *apsaras,* the celestial girls of Indian mythology.

Relief panels

On some of the temple mountains there are also relief panels illustrating various aspects of the royal mythology. Episodic relief sculpture first appears on Banteay Srei (10th century). The relief revolves around a series of Indian legends dealing with the cosmic mountain Meru as the source of all creation and with the divine origin of water. The chief artistic achievement of its architecture is the way in which it conceives and coordinates the spaces between the walls of the enclosures, the faces of the terraces, and the volumes of the shrine buildings. A most sophisticated architecture of full and empty space, it seems to have been influenced by that of the Hindu Pallava dynasty in southeastern India.

The earliest more or less complete example of a shrine complex devoted to deifying the ancestors of a king is the Preah Ko at Roluos, near Angkor, completed in 879. The earliest surviving temple mountain at Angkor itself is the Bakong, probably finished in 881. In the central shrine at the summit was a *linga,* the phallic emblem sacred to

Asie Photo—De Wys

Angkor Wat, Angkor, Cambodia, mid-12th century.

Śiva. Around the base of the terraced pyramid stood eight large shrines inside the main enclosure, with a series of moats, causeways, and auxiliary sculptures guarding the approaches to the exterior. The Bakheng, begun in 893, had an enormous series of 108 tower shrines arranged on the terraces around the central pyramid, which was crowned by a quincunx of principal shrines. The whole was intended to illustrate a mystical conception of the cosmos, very much on the lines of the great temple mountain at Borobudur in Java (see below *Indonesia*). Pre Rup, dedicated in 961, was probably the first of the temple mountains intended as a permanent shrine for the divine spirit of a king after his death. It, too, has a quincunx of principal shrines, but it is distinguished by the large number of auxiliary pavilions arranged along both sides of the inner enclosure wall.

From about the same period is perhaps the most beautiful—and most beautifully preserved—of the early Khmer temples, Banteai Srei. This was actually a private foundation, built some 12 miles from Angkor by a Brahmin of royal descent. Its auxiliary buildings, all of sandstone, are adorned with a profusion of elaborate ornament and relief figure sculpture. The roof gables, in particular, are treated with antefixes of fantastic invention. Its principal icon, a huge sandstone sculpture of the god Śiva, seated with his wife Umā on his left knee, is perhaps the most impressive full-round sculpture from the whole Khmer epoch. It differs from 10th-century Khmer official sculpture, which began to take on a conventional and relatively insensitive massiveness.

The Baphuon temple mountain (1050–66) is unfortunately almost completely destroyed. It was a vast monument 480 yards (440 metres) long and 140 yards (130 metres) wide, approached by a 200-yard (180-metre) causeway raised on pillars. Its ground plan shows that it was no mere assemblage of buildings but a fully articulated structure. In this it must rank as the immediate prototype for the great Angkor Wat. Built by Suryavarman II in the early 12th century, Angkor Wat is the crowning work of Khmer architecture, the culmination of all the features of earlier styles.

Angkor Wat

The enormous structure of the Wat is some 1,700 yards (1,550 metres) long by 1,500 yards (1,400 metres) wide. Surrounded by a vast external cloister, it is approached from the west by a magnificent road, which is built on a causeway and lined with colossal balustrades carved in the likeness of the cosmic serpent, associated with the sources of life-giving water. The Wat rises in three concentric enclosures. The western gate complex itself is nearly as large as the complex of central shrines, and both are subdivided into smaller, beautifully decorated courts. Only five of the original nine towers still stand at the summit; although they follow the basic pattern of the Khmer roof tower composed of diminishing imitative stories, the contour of the towers is not rectilinear but curved, so as to suggest that the stories grow one out of another like a sprouting shoot. All the courtyards, with their molded plinths, staircases, porticoes, and eaves moldings, are perfectly articulated enclosed spaces. The symbolic meaning of the Wat is clear. Its central shrine indicates the hub of the universe, while its surroundings—the gate complex, the cloister, the city of Angkor itself, and, finally, the whole visible world—represent the successive outer envelopes of cosmic reality. That it is oriented toward the west—and not to the east, as was customary—indicates that its builder, Suryavarman II, intended it as his own mortuary shrine; for, according to Indochinese mythology, the west is the direction in which the dead depart.

The sculptors who worked at the Wat demonstrate little ability in carving in the full round. Such full-round figures as there are—the guardians on the terraces, for example—lack life. The relief sculpture, however, is magnificent and full of vitality. The open colonnaded gallery on the first story contains over a mile of relief carving six feet (two metres) high. Much of it was originally painted and gilded, which strongly suggests that there must have been a Khmer style of painting of which nothing is known. The subject matter of the carvings is taken principally from the Hindu epics, but there are also many scenes representing Suryavarman's earthly glory. Working in relief only about an inch deep, the sculptors were able to depict an extraordinary complex of scenes of figures in vigorous action, full of complex overlaps to suggest deep space. The solid bodies are created mainly out of groups of convex curves; and everywhere there is the typical regional feeling for decorative spirals. Perhaps the most interesting group of figures are the *apsaras*, carved in relief, either singly or in groups, on the plain walls of the courtyards. These celestial beauties, whom Indian tradition describes as rewarding with their charms the kings, heroes, and saints who attain heaven, are carved with sinuous sensuality; but the most important part of their charm is their elaborate clothing, jewelry, and hairdressing or ornate, towering, jewelled crowns. Apparently, deep, downward-drooping curves standing far out from the body represented the height of Khmer chic. Skirts, stoles, and the long sidelocks of the hair all follow these curves, laid out flat on the ground of the relief. Symbolizing the erotic joys that are essential attributes of heaven, the *apsaras* were natural possessions of the king.

Bas-relief of a battle scene, Angkor Wat, Angkor, Cambodia, early 12th century.
Hotle Bildarchiv, Baden-Baden

The *apsaras*

In many senses the Wat was the end of the road for Khmer art. The effort demanded of the people in constructing this colossal stone monument, along with its four miles (six kilometres) of stone-lined moat 200 yards (180 metres) wide, appears to have been too great. The irrigation system itself may well have been neglected in favour not only of shifting the building stone—as much in quantity as there is in the Pyramid of Khafre in Egypt—but also of dressing, carving, and ornamenting it. After Suryavarman's death, the Cham, from the neighbouring kingdom of Champa (see below *Vietnam kingdom of Champa*), seized and sacked Angkor for the first time in its history (1177), thus shattering the confidence of the Khmer people in the protective powers of their Hindu deities. When Suryavarman's son, Jayavarman VII, came to the throne he inherited a ravaged kingdom. In 1181 he succeeded in driving out the Cham; he invaded their country and seized their capital, thereby making Champa a province of the Khmer. Then, over 60 years old, he embarked on a series of campaigns that extended the borders of the Khmer Empire further than ever before—into Malaya, Burma, and Annam.

The Bayon

The ruler of this empire naturally believed himself to be the greatest of the Khmers, and he set about demonstrating the truth of his belief by building his own city, Angkor Thom (*c.* 1200), and, at the centre of it, the biggest temple complex of them all—the Bayon (*c.* 1200). Breaking with all previous Khmer traditions, he took as his patron deity not one of the Hindu gods but one of the Buddhist *bodhisattvas*. Although Buddhism had flourished for several centuries in the whole of Indochina, it had not been

adopted by the Khmer as an imperial cult. Now that the Hindu gods had been discredited by defeat, Jayavarman placed himself under the patronage of Mahāyāna Buddhism. The mythology according to which the Bayon was designed was thus another version of the old mythology of the celestial mountain and the divine origin of water. Only the central figure of his mythology, Lokeśvara, Lord of the World, was specifically Buddhist. The colossal masks that look out over the four directions of the world from the towers of the Bayon and from the gates of Angkor Thom are there to demonstrate the compassionate, all-seeing power of Lokeśvara and the king.

When Jayavarman VII set out to create Angkor Thom, he had to raze the fine older work of his predecessors, for the site at Angkor had become choked with nearly four centuries of grandiose temple building. Within Angkor Thom's ten miles (16 kilometres) of moats he constructed huge complexes of building and made his city the focus of a final system of canals and irrigation, with additional lakes.

Unfortunately, the innumerable new shrines that surround Angkor Thom, the towers that crowd the Bayon, and the vast stone terraces faced with relief that surround the royal palace are in general much inferior in execution to the work of the earlier kings. Thus, today Angkor is dominated by the overwhelming presence of Jayavarman's immense but relatively unrefined architecture. The King's ambition was satisfied by size and quantity rather than artistic quality. Because sculptors were obliged to produce such vast quantities of work so fast, their standards deteriorated, and the huge vistas of narrative relief show signs of haste and slipshod workmanship. The real achievement lay with Jayavarman's scholastic architects, who conceived and laid out a complex of mythical imagery in massive architectural symbols. Their stupendous overall plan illustrates the creation of the world, a cosmos spreading outward from the central mountain tower. The two roads leading from the tower are lined with mile-long rows of gigantic deities who are pulling on the body of the serpent *nāga*. According to Hindu legend, the gods use the magical mountain Meru, symbolized by the mountain tower, as a churning stick and the body of the cosmic serpent as a churning rope to churn the world out of the milk of nothingness. Lake-sized fountains represent the healing waters of the Buddhist paradise, and allegories of salvation are realized in carved architecture. Perhaps the most impressive works of art associated with this last period of Angkor are some stone icons, such as the famous "Leper King," in the Angkor Thom complex. Many excellent smaller bronze figures of deities have also been found among the ruins.

Decline of Angkor

13th century to the present. After the death of Jayavarman VII, *c.* 1215, possibly as late as 1219, Angkor declined. The Thai population of Siam gradually pushed the Khmer down toward the Mekong Delta. Theravāda Buddhism became the religion of the people, and the grandiose vision of a cultural unity based on sacred kingship disappeared. In the 15th century, Angkor was retaken from the Thai, and a few buildings were restored by the ancestors of the modern (now abdicated) Cambodian kings. Some of the buildings were used as monasteries, but the city, with its essential irrigation system, had fallen into ruin.

Vietnam kingdom of Champa: c. 2nd to 15th century. The kingdom of Champa existed alongside the Khmer kingdom, sometimes passing under its rule, sometimes maintaining a precarious independence. From the north it was continually subject to the pressure of the advancing Vietnamese, a people racially related to the Burmese and Thai, who were themselves under pressure from the Chinese. The Hinduizing dynasties who ruled Champa from the 6th century were obliged to pay heavy tribute to the Chinese Empire. After 980 they were forced by the Vietnamese to abandon their northern sacred capital, My Son; thereafter, except for a brief return to My Son in the 11th century, their southern capital at Vijaya (Binh Dinh) became their centre. Under such disruptive circumstances, it is perhaps surprising that the Cham succeeded in creating and maintaining a dynastic art of their own. It was, however, always on a relatively modest scale, devoted to a conception of divine kingship similar to but far less ambitious than the Khmer.

The evolution of Cham art falls naturally into two epochs, the first when the capital was in the north, the second when it was removed to the south.

Art of the northern capital: 4th to 11th century. The form of the earliest temple at My Son, built by King Bhadravarman in the late 4th century, is not known. The earliest surviving fragments of art come from the second half of the 7th century, when the king was a descendant of the royal house at Chenla. The remains of the many dynastic temples built in My Son up until 980 follow a common pattern with only minor variations. It is a relatively simple one, with no attempt at the elaborate architecture of space evolved by the Khmer. Each tower shrine is based upon the central rectangular volume of the cell. The faces are marked by central porticoes that are blind on all but the western face, where the entrance door is situated. The blind porticoes seem to have contained figures of deities—perhaps armed guardians standing in a threatening posture. The porticoes are set in a tall, narrow frame of pilasters (columns projecting a third of their width or less from the wall), crowned with horizontally molded capitals that step out upward. They support a tall, double-ogival blind arch, crowned by another stepped in behind it. The arches are based on an Indian pattern and are carved with a design of slowly undulating foliage springing from the mouth of a monster whose head forms the apex of the arch. The faces of the walls are formed of pilasters framing tall recesses. The pilasters are carved with foliate relief, and elaborate recessed and stepped-out horizontal moldings mark their bases. The height of the pilasters and recesses gives a strong vertical accent to the body of the shrine. The principal architrave is carried on stepped-out false capitals to the pilasters. The roof of the tower is composed of three diminishing, compressed stories, each marked by little pavilions on the faces above the main porticoes. Inside the tower is a high space created by a simple corbel vault with its stepped courses of masonry. The chief portico was extended to include a porch, and the whole structure stood upon a plinth whose faces bore molded dwarfed columns (small columns) and recesses.

Pedestal altars

These temples have one distinguishing internal feature: a pedestal altar within the cell, upon which statues were set, sometimes, it seems, in groups. The pedestals themselves are often beautifully adorned with reliefs, and some of the best Cham sculpture appears upon them. The subjects are usually based on Indian imagery of the celestial court. The fact that the pedestal altars carried their sculptures in the space of the cell, away from the wall, meant that the Cham sculptors could think in terms of three-dimensional plasticity as well as relief.

The glory of Cham art is the sculpture of the whole of the first period. Much of what survives consists of lesser figures that formed part of an architectural decor: heads of monsters, for example, which decorated the corners of architraves, and figures of lions, which supported bases and plinths. These figures reflect the heavy ornateness of the Cham decorative style at its most aggressive; and many of them effloresce into the solid, wormlike ornament that is the Cham version of Indo-Khmer foliage carving and carries strong reminiscences of Dong Son work. The remaining fragments of the large icons suggest a double origin for Cham art traditions. On many of the capitals and altar pedestals are series of figures carved in relief in a sensuous style, which is nevertheless strictly conceptualized. This sophisticated work is reminiscent both of late Chenla art (see above *Cambodia and Vietnam*) and of Indonesian decoration, especially during the 11th-century return. Other figures are more coarsely emphatic in style, with the crudely defined cubic volumes and the heavy faces of Melanesian sculpture. It is thus probable that artists trained in the sophisticated Cambodian tradition worked for the Cham kings at one time or another, while Champa's own native craftsmen emulated the work of the foreigners in their own fashion.

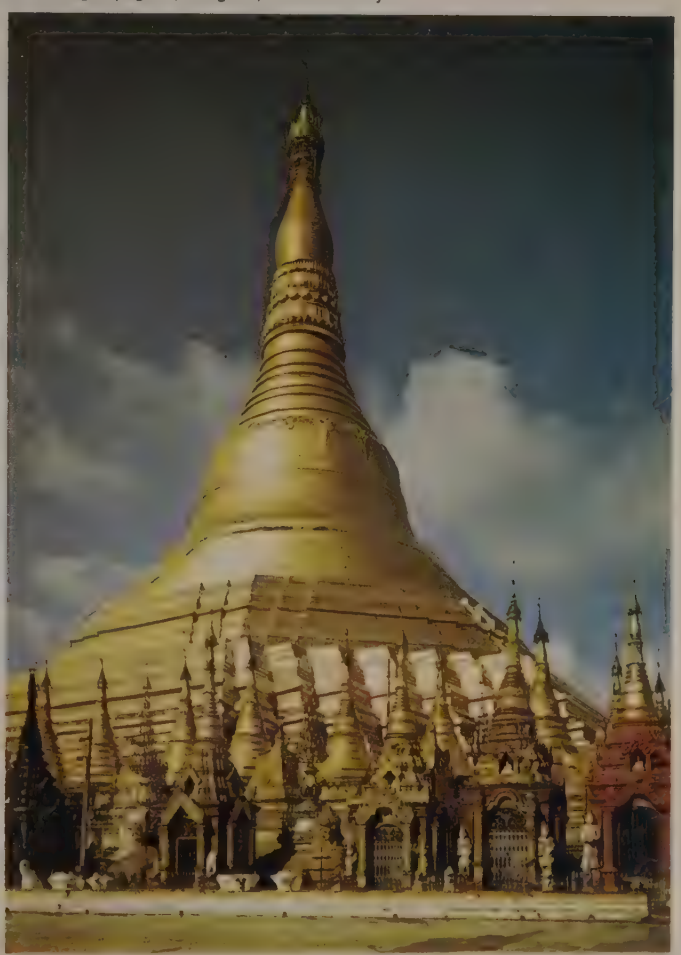
Apart from My Son there are one or two other sites in north and central Vietnam where Cham art was made in quantity. The most important of these is Dong Duong, in

Fresco of the Preaching Buddha at the Wet-kyi-in, Ky-byauk-kyi, Pagan, *c.* 1113.

Terra cotta relief at the Petleik-Matangjakata temple, Pagan, late 11th century.

Burmese art

Shwe Dagon pagoda, Rangoon, *c.* 15th century.



Vessel and cover in the shape of a sacred bird, gold decorated with filigree work and inlaid with rubies and imitation emeralds, 19th century. In the Victoria and Albert Museum, London. Height 41.5 cm.

Plate 1 By courtesy of (bottom left) the Victoria and Albert Museum, London photographs (top left, top right) J.A. Lavaud, Paris, (bottom left) A.C. Cooper Ltd., (bottom right) R. Manley from Shostal—EB Inc.

That Luang *stupa*, Vientiane, Laos, 1566, restored 18th and 19th centuries.

Bodhisattva from Nanchao, ancient Tai kingdom (now in the Chinese province of Yunnan), bronze, 13th century. In the British Museum. Height 44 cm.

Thai and Laotian art

Thai painted lacquer panel of a court scene, Bangkok style, middle of the 19th century. In the collection of Prince Piya Rangsit, Bangkok. Height 50 cm.

"The Great Departure of Bodhisattva," detail from "Episodes from the life of Buddha," Thai painting on silk panel, 17th-18th century. In the Musée Guimet, Paris. Detail 93 × 93 cm.

Plate 2: By courtesy of (top right) the trustees of the British Museum; photographs, (top left, bottom left) Hollie Bildarch v. Baden-Baden; (bottom right) Ciccione—Rapho—Photo Researchers

Plate 3: (Top left, centre) Hollie Bildarchiv, Baden-Baden, (top right) © Pascal Grelety-Bosviel, Paris (bottom) R. Manley from Susteal—EB Inc

Library, Banteay Srei, Angkor, Cambodia, 10th century.

"Siva and Umā," sandstone sculpture from Banteay Srei, Angkor, Cambodia, late 10th century. In the Musée National, Phnom Penh, Cambodia. Height 60 cm.

Cham sandstone panel of a pedestal altar showing an ascetic playing a flute, from Mi-son E 1, Vietnam, second half of the 7th century. In the Da Nang museum, Vietnam. Height 60 cm.

Cambodian and Vietnamese art

Gate at Angkor Thom, Angkor, Cambodia, *c.* 1200.

Pura Besakih, Gunung (mount) Agung, Bali, 14th century. In the centre is the eleven-storeyed *meru* dedicated to Siva.

Gold kris, embossed scabbard and grip, from southern Celebes. In the Royal Tropical Institute Museum, Amsterdam. Overall length 40.5 cm.

Indonesian art

Indonesian textiles. (Top) Javanese batik textile accented with gilding. In the Royal Tropical Institute Museum, Amsterdam. (Bottom) *Ikat* cloth from Sumba Timur, Lesser Sunda Islands. In the J. and R. Langewis Collection, Castricum, The Netherlands.

Main entrance to the Pura-desa Pasaban, Bali.

Plate 4: By courtesy of (centre left, centre right) the Royal Tropical Institute, Amsterdam; photographs, (top, bottom left, bottom right) Holle Bildarchiv, Baden-Baden

Quang Nam. It is a ruined Buddhist monastery complex of the late 9th century, conceived on the most beautifully elaborated plan of structured space in Champa. The architectural detail is distinguished from the My Son work by its greater emphasis upon the plasticity of architectural elements such as angle pilasters and porticoes. The circuit wall was about half a mile (one kilometre) long and once contained many shrines dedicated to Buddhist deities. It is possible that, when this complex of brick courts, halls, and gate pavilions was intact, it may have resembled very closely the contemporary Buddhist monasteries of northeastern India.

Art of the southern capital: 11th to 15th century. After 980, when the northern provinces were taken over by the Vietnamese and the Cham capital established at Binh Dinh in 1069, the kings maintained a gradually diminishing splendour. After the Khmer attack of 1145 they could claim little in the way of royal glory.

Although the Cham kings made a brief return to My Son from 1074 to 1080, most of their artistic effort was spent on shrines at Vijaya (Binh Dinh) and a few other sites in the south. The early 12th-century Silver Towers at Binh Dinh are simplified versions of the older northern towers, with corner pavilions added to the roofing stories and arches of pointed horseshoe shape. Throughout the 13th and early 14th centuries the architecture of successive shrines gradually declined. The plasticity of the old pilasters and architraves was suppressed into simple moldings, and the beauty of the buildings became largely a matter of proportion. By the mid-14th century even the temples erected at Binh Dinh amounted to little more than piles of crudely cut stones articulated only by reminiscences of the classic Cham style.

Decline of shrine architecture

Sculpture shows a parallel decline. One or two reliefs at the Silver Towers do convey a sense of tranquillity and splendour, but an indigenous style of rigid cubical emphasis came progressively to dominate the iconic Hindu figures at southern sites. The curlicued design of earlier figures was gradually converted into a style of massive, scarcely carved blocks that convey, at their best, an impression of barbarous strength but without the refinement of first-class primitive art.

As was the case in Cambodia, this decline in art by the mid-14th century may be attributed to the people's loss of confidence in the concept—and, with it, the imagery—of divine kingship. Theravāda Buddhism, as a popular religion based upon numerous small, local monasteries, adopted probably from the Tai, was spreading all over the region. The northern Vietnamese, who had originally been organized in self-contained kingdoms without any concept of royal divinity, owing an intermittent administrative allegiance only to the distant Chinese emperor, found this ultimately suitable as a state religion after the final eclipse of Confucianism in the 17th century. They did incorporate echoes of older Hindu architecture, however, in details of the flamboyant ornament used on eaves and gables of their wooden monastery buildings.

Vietnam: 2nd–19th century. The great achievement of Vietnamese art, at least during the Le period (15th–18th centuries), seems to have been in architectural planning, incorporating Confucian, Taoist, or Buddhist temples into the landscape environment. The plans themselves include halls for a multitude of images in South Chinese vein and provision for a variety of rituals. There are no intact monuments of early Vietnamese architecture that are unrestored. Numerous fragments exist, however—either isolated stone bases, columns, stairways, and bridges or carved wooden members incorporated into later buildings—all of which are influenced to some degree by Chinese styles.

Tombs of the Chinese type

Tombs of generically Chinese type from the 2nd to 7th century contain bronze furnishings, in many of which, such as lampstands, the influence of the Dong Son style is clearly visible. There are no spirit images so typical of Six Dynasties (3rd–6th centuries) and T'ang (7th–10th centuries) Chinese tombs. The Chua Mot-cot, Hanoi, has vestiges of a stone shrine probably dated 1049. The only old paintings, on rock, at Tuyen Quang, (9th century) represent the Buddha, *bodhisattvas*, and donors. The Van-mieu at Hanoi (built 1070 but frequently restored) contains ritual bronzes in "barbaric" Chinese style.

Perhaps the most interesting early sculptures to survive are the stone fragments from the Van-phuc temple (9th–11th centuries), which are based on Chinese Buddhist imagery but in a style strongly Indianized, perhaps by Cham influence. The most important piece of old work still virtually intact is the portable octagonal wooden stupa kept in the hall of the But-thap, at Bac Ninh, east of Hanoi. It has wooden panels carved in a flamboyant 14th-century Chinese style; part of it bears a representation of the Buddhist paradise of Amitābha. Incorporated in many Buddhist temples of the Le period (15th–18th centuries), as well as in stone terraces, bridges, and gateways, is extremely elaborate carved and coloured woodwork in a style based upon the coiling dragon-and-cloud decoration of Ming (1368–1644) and Ch'ing (1644–1911) China, but with a characteristically Vietnamese exaggeration of weight and curve.

At Tho Ha there was a potters' village, where the glazed ceramic figures used on many types of Chinse temple were manufactured. The remains of many tombs, palaces, bridges, and Confucian and Taoist temples decorated in similar vein are known everywhere.

19th and 20th centuries. The beautiful imperial palace of Hue (final plan before 1810) contained vestiges of older architecture and many works of Sinicized art before its devastation in 1968. It consisted of a series of simple, rectangular, one-story pavilions, laid out among trees inside a group of courts. These buildings and their decoration were southern Chinese in basic conception.

Elsewhere in Vietnam, both religious and secular buildings have been constructed in the 20th century in provincial versions of Chinese styles. There was little demand for the sculptor's art beyond the carving of stereotyped Buddha icons, monsters, and guardians. In modern times southern Vietnam has adopted a decorative style partly derived from the active traditions of Bangkok. Religious sects abound, with hybrid native European and Chinese elements used in their iconic and decorative art. Because of political turmoil, no clear and individual modern Vietnamese artistic tradition has been able to emerge.

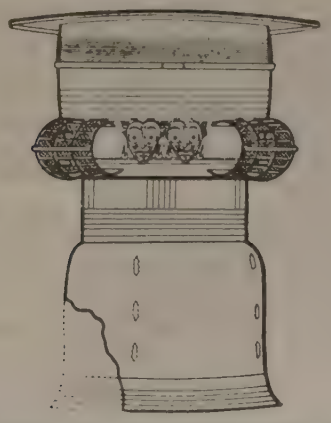
INDONESIA

The islands that at the present day compose Indonesia probably once shared in the complex Neolithic heritage of artistic tradition, which also spread further, into the islands of Melanesia and Micronesia. Beautifully ground Neolithic axes of semiprecious stone are treasured still in some countries. In many parts of Indonesia there are quantities of megalithic monuments—menhirs, dolmens, terraced burial mounds, stone skull troughs, and other objects. Some of these are undoubtedly of Neolithic date, but megaliths continued to be made in much more recent times. One stone sarcophagus in eastern Java, for example, is dated post-9th century. On Nias island, megaliths are still revered, and they are still being erected on Sumba and Flores islands. Thus, in Indonesia especially, different layers of Southeast Asian culture have existed side by side. The most impressive and important collection of megaliths is in the Pasemah region, in south Sumatra, where there are also many large stones roughly carved into the shape of animals, such as the buffalo and elephant, and human figures—some with swords, helmets, and ornaments and some apparently carrying drums.

Megalithic monuments

These drums immediately suggest the drums characteristic of the mainland Southeast Asian Dong Son culture, which flourished *c.* 4th–1st centuries BC (see above *General development of Southeast Asian art*). This culture may well have helped to diffuse throughout the region styles related to Chinese Chou and pre-Han ornamental work. Certainly, the Dong Son influence is clear in many of the ceremonial axes, as well as many of the ornamented bronze drums, that have been found in the islands. The bronzes were cast by a cire perdue process resembling that used in parts of the Asian mainland. The largest and most famous drum is "the Moon of Bali," found on that island near Pedjeng. It has molded flanges, and cast onto its faces is extremely elaborate relief ornament consisting

of stylized masks with ears pierced and lengthened by large earrings. Such drums were probably originally used in ritual—by the rainmaker, perhaps—and they may have been buried with the distinguished dead. No one knows the exact age of these bronzes; "the Moon of Bali," for example, could be anywhere between 1,000 and 2,000 years old. Similar small drums were used quite recently as bride prices; and many of the islands today produce textile designs and ceremonial bronzes that are strikingly reminiscent of Dong Son ornament.



"The Moon of Bali," bronze drum found near Pedjeng, Bali. In the Panataram Sasih temple, Pedjeng, Bali. Height 1.87 m, diameter 1.60 m.

From P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

Central Javanese period: 7th to 13th century. Sometime between the 3rd and 6th centuries AD, Indianized principalities existed in Java. The chieftains who lived in their *kratons* (fortified villages) seem to have derived great inspiration, prestige, and practical assistance from the skills and ideas imported from India. In Sumatra there was the important but so far enigmatic Indianized kingdom of Śrivijaya, which, from its strategic position on the Strait of Malacca, exercised a powerful artistic influence in the whole region. Its great Buddhist centre, Palembang, might have had direct connections with the monasteries of southeastern India; for fine bronze Buddhas and *bodhisattvas* in a style reminiscent of Amarāvatī (2nd century AD) have been found in many regions where the influence of Śrivijaya might have been felt, including Mon Dvaravati (see above *Thailand and Laos*) and distant Celebes. Elsewhere among the islands were Indianized kingdoms still unknown to history.

The Indianized art of Śrivijaya

The local dynasties of the *kratons* competed among themselves for power, and eventually the principal dynasties known to history came to the fore. The earliest major cultural assimilations from India took place probably during the 7th century, when the Hindu Pallava form of southeast Indian script was adopted for inscriptions in west Java. Thereafter, a central Javanese dynasty that worshipped Śiva made the oldest surviving artworks in stone. The last king of this dynasty retreated to east Java in the face of the rising power of another central Javanese dynasty, the Śailendra (AD 775–864.) The Śailendra were followers of Mahāyāna and Tantric forms of Buddhism, although Hinduism, as manifested in the worship of Śiva and Vishnu, was by no means eliminated. This dynasty created far the larger part of the immense wealth of first-class art known today in Java.

In Indonesia, the word *tjandi* refers to any religious structure based on an Indianized shrine with a pyramidal tower. This was the essential form on which virtually all the stone Indianizing architecture of Southeast Asia was originally based. The Javanese, like the Khmer, evolved an elaborate architecture of their own around the basic Indian prototype.

Central Javanese stone architecture did not use structural pillars, nor did its major stone monuments conceptualize hollow space in the way Khmer architecture did. Like Indian stonework, central Javanese stonework is fundamentally conceived as a solid mass, serving as a vehicle for figurative and symbolic sculpture. Its temples are centralized, with enclosures radiating around the central shrine. In eastern Java and Bali, however, the pattern of the shrine was influenced by older traditions and was usually conceived as an enclosure, the walled area of ground be-

By courtesy of (centre, right) the Royal Tropical Institute, Amsterdam, photograph, (left) Louis Frederic—Rapho/ Photo Researchers

Indonesian sculpture.
(Left) "Buddha" from Celebes, bronze, Amarāvatī style, 3rd–5th century. In the Jakarta Museum. Height 75 cm. (Centre) "Vishnu" from Bali, stone. In the Royal Tropical Institute Museum, Amsterdam. Height 80 cm. (Right) Ancestor figure from the Tanimbar Islands, Indonesia. In the Royal Tropical Institute Museum, Amsterdam. Height 38 cm.

ing the sacred element, while the buildings in it were of secondary importance. Old wooden buildings do not survive; but representations of wooden architecture in stone reliefs and the recent architecture of Bali show that eastern Indonesia was influenced by the ancient Southeast Asian tradition of constructing wooden pillared halls with tiered, sloping, and gabled roofs.

Because there are no inscriptions to supply dating points, the exact dates of the earliest Indonesian architectural monuments are not certain. The group of shrines generally believed to be the earliest is situated on the Dijeng Plateau. This is a high volcanic region, about 6,000 feet (2,000 metres) above sea level, where there are sulfur springs and lakes. The whole mountain seems to have been sacred to the Hindu deity Śiva, for all temples on the Dijeng are dedicated to him. There can be little doubt that during the 8th and 9th centuries the Javanese, who traditionally had interpreted the volcanic turbulence of their landscape as a manifestation of divine power, identified this power with the terrifying Śiva. On other Javanese volcanic mountains, also, groups of shrines are dedicated to him.

Hindu *tjandi*s

The temples on the Dijeng are single-cell shrines, roofed with diminishing stories. The exteriors of the temples are relatively plain; only around door frames and window frames are there distinctive passages of central Javanese ornament. Around the niches of Tjandi Puntadewa are perhaps the earliest surviving examples of the characteristic Javanese doorframe: across its lintel is carved a mask of the Indian Kāla monster, which represents time; and down the jambs, as if vomited from his open mouth, run string panels of foliage. The foot of each jamb terminates in an elaborately carved scrollwork cartouche, which is itself a *makara* (water monster) head seen in profile. This *tjandi*, like others on the Dijeng, has a single approach stairway rising between curved balusters. A few stone images of Śiva from these temples have been found. In broad, vigorous forms they express the dangerous power of the god.

Two of the very finest early Javanese sculptures—virtually in the full round—come from yet another Śiva temple, Chandi Banon, near Borobuḍur (see below). One, representing the god Vishnu (no stranger in syncretic Javanese temples of Śiva), has the extremely smooth, faintly amorphous suavity, the absolute convexity, and the lack of definition between planes characteristic of the classical central Javanese sculptural style; while the garment he wears, with its assortment of girdles, is closely reminiscent of late Pallava–early Cōḻa Hindu styles of southeast India. Another icon, sometimes called Agastya but more likely the third deity of the Hindu trinity, Brahmā, represents the god in the form of a bearded Brahmin sage. He has a large and, to Oriental eyes, splendid potbelly. This icon was indigenous to southeast India. The great depth of the side recessions of these figures, although perhaps not so clearly defined as in the great Funan–Chenla style (see above *Cambodia and Vietnam*), gives them a bland massiveness. The lack of movement in the figures and the regularity of the designs, the impassive faces, and the slowness of the lines must have been part of the central Javanese conception of transcendent glory.

The Hindu temples of central Java are conceived simply as shrines to contain icons of deities for worship. The Mahāyāna and especially the Tantric Buddhist *tjandi*s, however, were called upon to do far more. They were designed to express complex metaphysical theories. The challenge this presented to the central Javanese architects was met in a series of splendid monuments, completely original in conception. The culminating work of the series, Borobuḍur, is a highly evolved architectural image, whose subtlety and refinement were never matched, even at Angkor in Cambodia.

Buddhist *tjandi*s

The first work of this Buddhist series is Tjandi Ngawen, near Muntilan. This *tjandi* consists of five shrines facing east, 12 feet (four metres) apart in a row from north to south. Each shrine contained one of the five Buddhas who, according to Tantric Buddhist theory, presides over one of the five major psychological categories under which ultimate reality reveals itself. The shrines themselves are based on but more developed than those used for Hindu deities elsewhere in Java. Roughly square in plan and roofed with diminishing stories, they have pilastered projections on three faces and a portico on the east. Along the architrave are small triangular antefixes and reliefs of Kāla monsters vomiting floral scrolls hood the niches and portals.

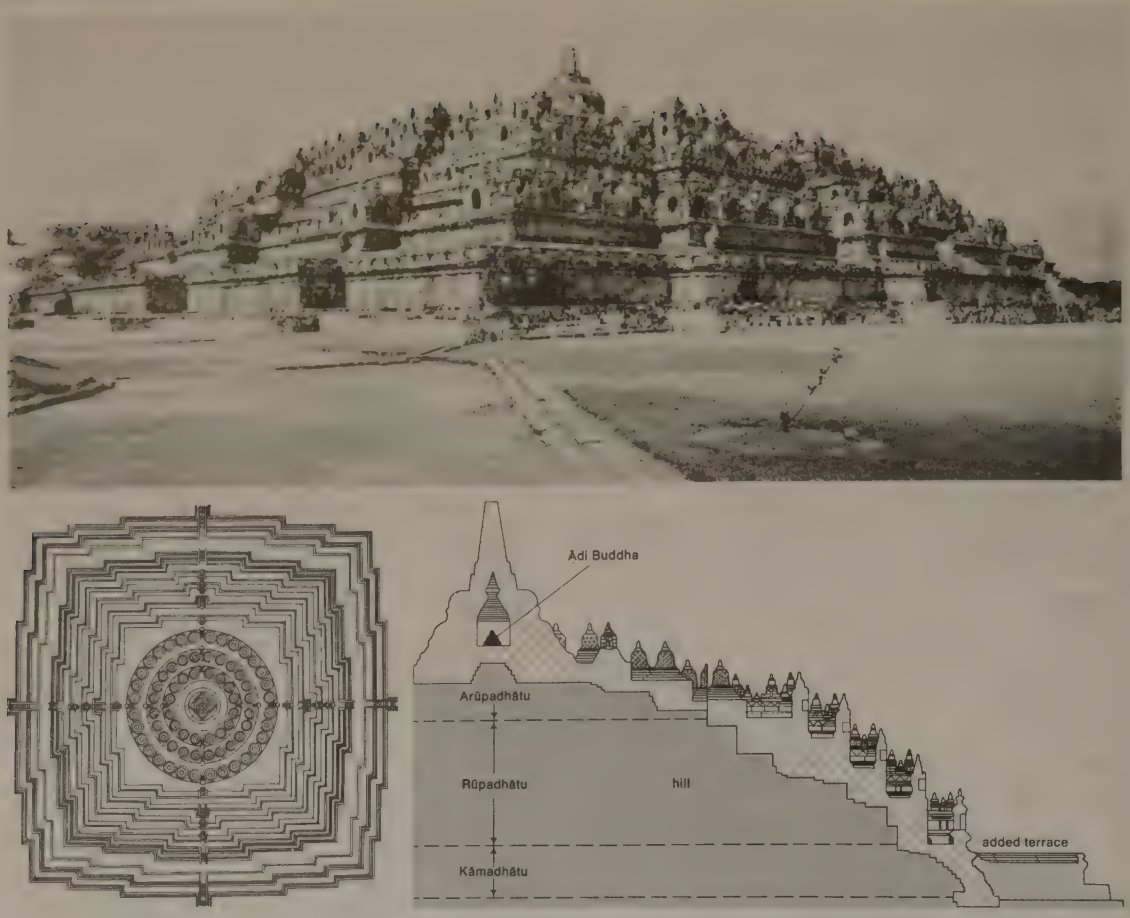
The group of five Buddhas is familiar in the art of Tibet, Japan, and northeast India. Among them they compose what is called the *vajra-dhātu*, which means, roughly speaking, "the realm of total reality." According to the old Javanese theology, above this group is another, called the deities of the *garbha-dhātu*. *Garbha* means "womb" or "innermost secret," and its three deities personify the most esoteric realms of Buddhist speculation. At the centre of the group is the image of the single, undivided Buddha nature, which symbolizes the ultimate reality of the entire universe. From his right side emanates the *bodhisattva* Lokeśvara (Lord of the World), who is both compassionate and possessed of all power. From the left emanates the *bodhisattva* Vajrapāṇi, who is the personification of the most secret doctrines and practices of Tantric Buddhism. One of Java's greatest monuments, Tjandi Mendut, is a shrine expressly created to illustrate the combined doctrine of *garbha-dhātu* and *vajra-dhātu*.

Mendut dates from about 800 and is thus, generally speaking, contemporary with Borobuḍur. It is formed as a single large, square chamber, roofed with the usual diminishing stories, and mounted on a high, broad plinth.

Tjandi Mendut

Louis Frederic—Rapho/Photo Researchers

Tjandi Mendut, near Borobuḍur, Java, *c.* 800.



(Top) Borobuḍur, Java, *c.* 800. (Bottom left) Plan of Borobuḍur and (bottom right) section of Borobuḍur.
(Top) Holle Bildarchiv, Baden-Baden, (bottom left, bottom right) from P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

which is approached on its northwestern face by a staircase with recurved balustrades. The exterior is in every way more ornate than that of any shrine so far discussed. In addition to floral diaper (an allover pattern consisting of one or more small repeated units of design connecting with or growing out of one another) and scrolls, there are numerous figures in relief representing male and female deities, the subsidiary principles of the combined doctrine of *garbha-dhātu* and Vajrapāṇi. Cut into the fine ashlar (squared-stone) masonry are many relief panels with scenes from Buddhist literature, each panel self-contained and placed with consummate aesthetic judgment. Some represent mythical ideas, such as the wish-granting tree, others narratives from Buddhist legend.

The principal images were placed inside the cell chamber. Apparently, there were originally seven huge stone icons, but only three remain: the central Buddha, who also represented the ultimate Buddha nature of the *garbha-dhātu,* and his two emanations in the *garbha-dhātu,* Lokeśvara, and Vajrapāṇi. When complete, the interior of Mendut must have been an even more awe-inspiring and spiritually moving place than it is now. The three great statues are seated on elaborate thrones, backed against walls, but the figures are carved virtually in the full round. The inflated, gently inflected forms of the figures give them a majestic presence. The types and carving technique, as well as the monumental scale of the figures, are reminiscent of contemporary work in the cave temples of the western Deccan in India.

On the west–east road from Tjandi Mendut to Borobuḍur stands a small, relatively plain temple called Tjandi Pawon, dedicated to the god of wealth. Pawon was probably a kind of anteroom to Borobuḍur, catering to the more worldly interest of pilgrims. The outside has fine reliefs of female figures, and the roof bears towers of small stupas. On the reliefs are wish-granting trees surrounded by pots of money, and bearded dwarfs over the entrance pour out jewels from sacks.

Borobuḍur

Borobuḍur is one of the most impressive monuments ever created by man. It is both a temple and a complete exposition of doctrine, designed as a whole, and completed as it was designed, with only one major afterthought. It seems to have provided a pattern for Hindu temple mountains at Angkor (see above *Cambodia and Vietnam*), and in its own day it must have been one of the wonders of the Asian world. Built about 800, it probably fell into neglect by about 1000 and was overgrown. It was excavated and restored by the Dutch between 1907 and 1911. It now appears as a large, square plinth (the processional path) upon which stand five terraces gradually diminishing in size. The plans of the squares are stepped out twice to a central projection. Above the fifth terrace stands a series of three diminishing circular terraces carrying small stupas, crowned at the centre of the summit by a large, circular, bell-shaped stupa. Running up the centre of each face is a long staircase; all four are given equal importance. There are no internal cell shrines, and the terraces are solid; Borobuḍur is thus a Buddhist stupa in the Indian sense. Each of the square terraces is enclosed in a high wall with pavilions and niches along the whole perimeter, which prevents the visitor on one level from seeing into any of the other levels. All of these terraces are lined with relief sculptures, and the niches contain Buddha figures. The top three circular terraces are open and unwalled, and the 72 lesser, bell-shaped stupas they support are of open stone latticework; inside each was a huge stone Buddha figure. The convex contour of the whole monument is steepest near the ground, flattening as it reaches the summit. The bottom plinth, the processional path, was the major afterthought. It consists of a massive heap of stone pressed up against the original bottom story of the designed structure, so that it obscures an entire series of reliefs—a few of which have been uncovered in modern

Sculpture at Borobuḍur.
(Left) Stupas on one of the circular terraces. The figure in the stupa in the foreground is a Buddha. (Right) Bas-relief on the exterior wall of one of the circular terraces.
By courtesy of (right) the Indonesian Tourist Board, photograph (left) Josephine Powell, Rome

times. It was probably added to hold together the bottom story, which began to spread under the pressure of the immense weight of earth and stone accumulated above.

Symbolism of Borobuḍur

The whole building symbolizes a Buddhist transition from the lowest manifestations of reality at the base, through a series of regions representing psychological states, toward the ultimate condition of spiritual enlightenment at the summit. The unity of the monument effectively proclaims the unity of the cosmos permeated by the light of truth. The visitor was meant to be transformed as he climbed through the levels of Borobuḍur, encountering illustrations of progressively more profound doctrines the nearer he came to the summit. The topmost terrace, whose main stupa contained an unfinished image of Buddha that was hidden from the spectator's view, symbolized the indefinable ultimate spiritual state. The 72 openwork stupas on the circular terraces, with their barely visible internal Buddhas, symbolize incomplete states of enlightenment on the borders of manifestation. The usual way for a pilgrim to pay reverence to a Buddhist stupa is to walk around it, keeping it on his right hand. The vast series of reliefs about three feet (one metre) high on the exterior walls of the terraces would thus be read by the visitor in series from right to left. Between the reliefs are decorative scroll panels, and a hundred monster-head waterspouts carry off the tropical rainwater. The gates on the stairways between terraces are of the standard Indonesian type, with the face of the Kāla monster at the apex, vomiting his scrolls.

The reliefs of the lowest level illustrate scenes that show the causal workings of good and bad deeds through successive reincarnations. They show, for example, how those who hunt, kill, and cook living creatures such as tortoises and fish are themselves cooked in hells or die as children in their next life. They show how foolish people waste their time at entertainments. From these scenes of everyday life, one moves to the terraces above, where the subject matter becomes more profound and metaphysical. It illustrates important Mahāyāna texts dealing with the self-discovery and education of the *bodhisattva,* conceived as being possessed by compassion for and devoted wholly to the salvation of all creatures. The reliefs on the uppermost terraces gradually become more static. The sensuous roundness of the forms of the figures is not abated; but, in the design, great emphasis is laid upon horizontals and verticals and upon static, formal enclosures of repeated figures and gestures. At the summit all movement disappears, and the design is entirely subordinated to the circle enclosing the stupa.

The iconography of Borobuḍur suggests that the legend of the royal *bodhisattva* recounted in many of the reliefs was meant to "authenticate" some king or dynasty. Yet it hardly seems possible that Borobuḍur was the focus of a specific royal cult, as there is no provision at all for the performance of royal ritual. It must have been, then, in some sense a monument for the whole people, the focus for their religion and life, and a perpetual reminder of the doctrines of their religion.

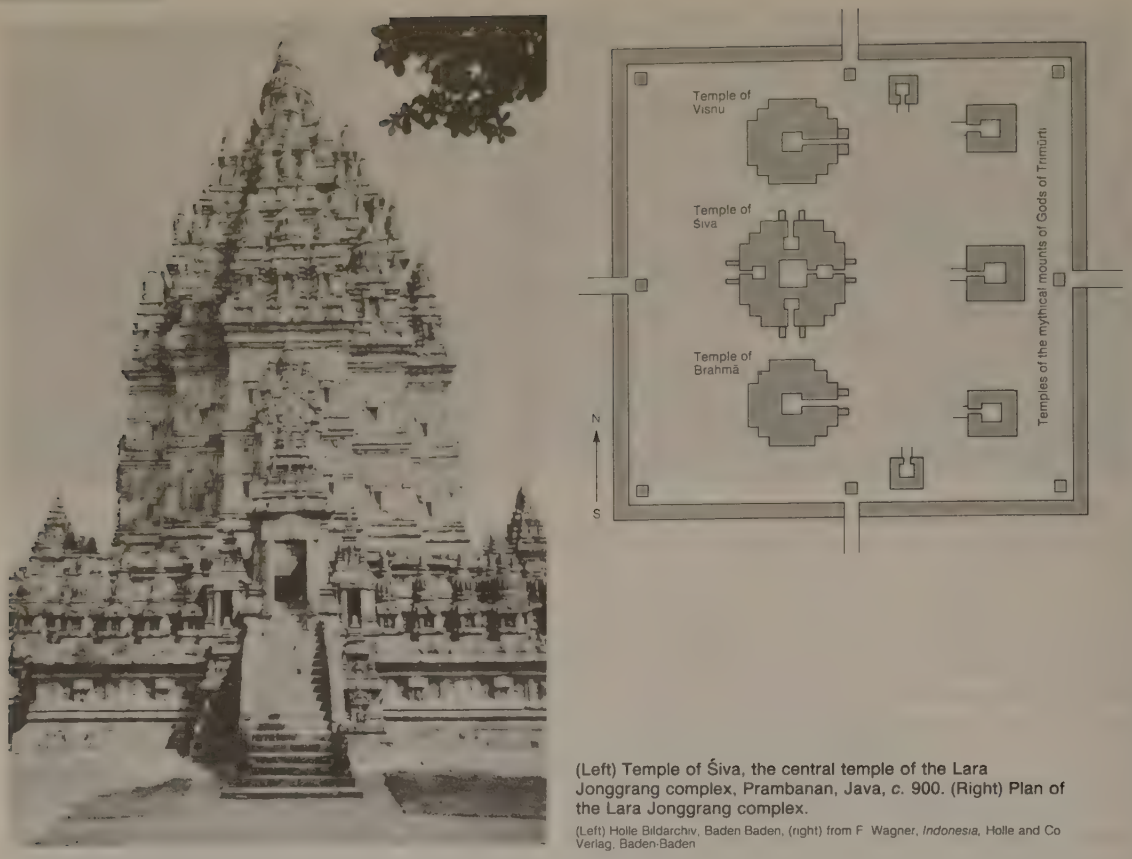
A considerable number of bronzes, some small, some large, have been found in Indonesia in a style close to that of the sculptures of Borobuḍur and Mendut. One fine, large standing image comes from Kotabangun in Borneo; but some come from Java. Many small cult images of the Buddha and Buddhist deities exist. Some are close in type to the early Pāla images of Indian Bihār, the homeland of Buddhism, with which the Javanese must have maintained close touch. A few small but extremely fine gold figurines of undoubted Javanese workmanship have also turned up. For all their small size they must rate as first-class works of art. As well as images there are many beautiful bronze ceremonial objects, such as lamps, trays, and bells. These objects are decorated with the same kinds of ornament, although on a miniature scale, as the architectural monuments: scrolled leaves, swags, and bands of jewels.

Post-Borobuḍur *tjandi*s illustrate the Buddhist doctrine in different ways. Kalasan, for example, built in the second half of the 8th century, was a large, square shrine on a plinth, with projecting porticoes at the centre of each face. The roof was surmounted by a high circular stupa mounted on an octagonal drum, the faces of which bear reliefs of divinities. Topping each portico was a group of five small stupas, and another large stupa stood at each disengaged corner of the main shrine. The moldings were restrained and elegantly profiled. Each section of the exterior wall contains a niche meant for a figure sculpture. The decorative scroll carving is especially fine.

Another shrine from this period, Tjandi Sewu, consisted of a large cruciform shrine surrounded by smaller temples, only one of which has been restored. All of the temples seem to have had roofs in the form of tiered stupas, compressing the overall Borobuḍur scheme into the scope of a storied shrine tower. From Tjandi Plaosan came many beautiful sculptures, donor figures, and iconic images of *bodhisattva*s.

Tjandi Sari

Perhaps the most interesting of the post-Borobuḍur Buddhist shrines of the 9th century is Tjandi Sari. It is an outstanding architectural invention. From the outside it appears as a large, rectangular, three-storied block, with the main entrance piercing the centre of one of the longer sides. The third story stands above a substantial architrave with horizontal moldings and antefixes. Two windows on each short side, three on each long, open into each story, though at the rear they are blind. The windows are crowned by large antefix-like cartouches of ornamental carving based on curvilinear pavilions hung with strings of gems. The uppermost windows are hooded with the Kāla-monster motif. The roof bears rows of small stupas, and perhaps there was once a large central stupa. Inside, Tjandi Sari contains a processional corridor around three



(Left) Temple of Śiva, the central temple of the Lara Jonggrang complex, Prambanan, Java, *c.* 900. (Right) Plan of the Lara Jonggrang complex.

(Left) Holle Bildarchiv, Baden Baden, (right) from F Wagner, *Indonesia*, Holle and Co Verlag, Baden-Baden

interior shrines that were possibly intended for images of the *garbha-dhātu* deities, as at Tjandi Mendut.

The last great monument of the central Javanese period, Lara Jonggrang at Prambanan, is indeed a colossal work, rivalling Borobuḍur. It was probably built soon after 900. Not Buddhist but Hindu, the shrine represents the cosmic mountain. There were originally 232 temples incorporated into the design. The plan was centred on a square court with four gates containing the eight principal temples. Facing east, the central and largest temple, some 120 feet (40 metres) high, was devoted to the image of Śiva. To the north and south it is flanked by slightly smaller temples devoted to the two other members of the Hindu trinity, Vishnu and Brahmā. The smaller shrines contained many subsidiary images. The whole complex was enclosed, far off-centre, in an extremely large walled courtyard.

Although these are Hindu buildings, their high-terraced shrine roofs bear tiers of elongated and gadrooned stupas. The reliefs on these structures are especially beautiful. One series, representing the guardians of the directions, integrates the ornamental motifs with the plastic forms of the bodies in a most original way. The balustrades and inset panels abound with lively reliefs portraying various deities or scenes taken from the great Hindu classics, especially the *Rāmāyaṇa*.

East Javanese period: 927–16th century. During the east Javanese period a very large number of monuments were produced at the eastern end of the island (after 1222) and in Bali (after *c.* 1050). Few single structures, however, are as impressive and as comprehensively planned as are the monuments of Borobuḍur or Lara Jonggrang.

Around the strange natural mountain with tiered peaks cut and built in stone called Mount Penanggungan there are 81 structures (10th-century) of different kinds (now mostly in ruins). Prominent among these structures are bathing places. This mountain was identified by the people with the sacred Mount Meru, and its natural springs were believed to have a magical healing power and a mystical purifying capacity. Another such bathing place is Belahan (11th century). Made of brick, it, too, has extensive ruined temples. Belahan is supposed to have been the burial place of King Airlangga, who probably died about 1049. One of the greatest east Javanese icons formed the central figure against the back wall of the tank. Carved of red tufa (a porous rock), it shows the god Vishnu seated at peace on the back of his violently dramatic bird-vehicle, Garuḍa. It is said that the image represents the King himself in divine guise. Beside this image was a sculpture of a type associated with many of these sacred bathing sites. It is a relief of a four-armed goddess of abundance, her two lower hands holding jars pierced with holes, her two upper hands squeezing her breasts, which are also pierced; through the holes the sacred water flowed into the basin. There are many variants of this idea at the springs of Mount Penanggungan. On Bali the same kind of fountain sculpture appears at the Goa Gadjah, at Bedulu, in a spring-fed tank below a cave.

Four-armed goddess of abundance

In both Java and Bali there are many rock-face relief carvings from this period (there are no secure dates). Some represent legendary scenes; others represent *tjandis*; the shallow chambers of others are thought to be royal tombs.

The structure that gives the best ideas of what the typical east Javanese shrine of the mid-13th century was like is Tjandi Kidal. The nucleus of the building is a square cell, with slightly projecting porticoes each hooded by an enormous Kāla-monster head. But the cell itself is dwarfed both by the massive molded plinth upon which it stands and by the huge tower with which it is surmounted. The tower stands above an architrave stepped far out on tiered moldings. It is no longer composed of diminishing stories, as earlier towers were, but is conceived as a massive pyramidal obelisk made up of double bands of ornament spaced by stumpy pilasters and bands of recessed panels. The architectural projections and moldings distinguish Tjandi Kidal from earlier Javanese architecture, with its plain wall surfaces.

Many masterpieces of sculpture belong to the east Javanese period. Among them are some superb icons of Śiva and of a goddess of Buddhist wisdom from Singhasāri and a splendidly "primitivist" image of the elephant-headed god of wealth from Bara, Blitar.

From the late 13th century onward a whole series of *tjandis* was created in eastern Java. As time went on the *tjandis* lost their monumental scale and became simply

shrines within a series of courtyards on a pre-Indian pattern. From Tjandi Djago through Tjandi Panataran at Blitar (14th century) and Tjandi Surawana it is possible to trace the line of descent of the modern Balinese temple enclosures.

Influence of shadow puppets on relief figures

By the end of the 14th century, the figures in the relief sculpture at these shrines had come more and more to resemble the shadow puppets of the popular *wayang* drama. They adopt the stiff profile stance that presents both shoulders, while the trees and houses resemble the stereotype silhouette leather and wood cutouts used as properties in the shadow plays. The art of carving in the near-full round, however, did not follow the same course of evolution as the reliefs. Such work did become softer and more delicate in style, with accretions of broad floral forms, but well into the 15th century the icons retain something of the strength of older sculptural conceptions. Another plastic tradition that seems to have escaped domination by the *wayang* formula resulted in the production of beautiful small terra-cotta figures as part of the revetment (stone facing sustaining the embankment) of the east Javanese capital city of Majapahit. Like the reliefs, the many small excavated bronzes of Hindu scenes are under the *wayang* influence, three-dimensional though they may be. Curlicues proliferate, and the plasticity of bodies is virtually ignored.

16th century to the present. When Islām arrived in Indonesia, it used the repertoire of traditional ornament for its mosques and tombs; but, in conformity to a puritan Muslim custom, the representation of living creatures was excluded on religious buildings. The gates of the 16th-century mosque at Sendangduwur, Badjanegara, show a splendid example of this adaptation. The wings of the old Hindu Garuḍa, a colossal bird-vehicle of the high god Vishnu, frame the gate; the body and head are suppressed. Above the lintel are abstract tree-clad mountain forms recalling the imagery of the cosmic Meru; and legendary snakes hood the jambs. The 16th-century mosque at Kudus even has a gate based on the split-*tjandi* pattern used in Bali (see below). Tombs such as that of Ratu Ibu at Airmata, on the island of Madura, add to their simple volumes elaborate but abstract variants of the scroll-filled antefixes of older architecture and of the petal-shaped aureoles of the larger east Javanese icons. In Sumatra the Muslim rulers encouraged a revival of the pre-Indian ancestor cult, along with its ancient and characteristic arts.

Bali. The rajas of eastern Java finally retreated before the Muslim invaders during the 16th century and departed to the island of Bali, where they remained. The old Javanese Indianized culture they brought with them survived and combined with animist folk elements. In Bali today that culture has bred a widespread popular art. There are now many hundreds of temples in Bali of varying age. Each family group has its own temple, dedicated to the ancestors; each village, too, has its temple, in which special attention is paid to a rich fertility goddess identified with the ancient Indian goddess of bounty, Śrī. Special temples dedicated to the goddess of death stand near the cremation ground. There are numerous major temples—many associated with volcanic peaks—dedicated to different deities and spirits; they range in size and importance from Besakih on Mount Agung (where a megalith is incorporated as a phallic Śiva-emblem) to Panataram Sasih of Pedjeng (where the bronze drum called "the Moon of Bali" is preserved).

Balinese temples

Balinese temples are conceived as multiple courts raised on terraces. The tall stone or brick and plaster gates are shaped like a *tjandi*-tower split down the centre; they are usually encrusted with ornament based upon deep multiple curlicues interspersed with simplified, two-dimensional relief figure sculpture. Fantastic three-dimensional guardians sometimes stand at the foot of the access staircase. Beyond the gates are one or two courts within which various ceremonies (including sacrifices and cockfights) may take place. The rearmost court backs onto the mountain, whence spirits descend temporarily when invoked. The court has no icons; at most, there is a seat for invisible deities. The structures in the court, mostly of wood and thatch, may be of many stories. (Such structures are called *merus*.) Sometimes the treasuries are ornamented with carving; and a few older stone *meru* towers in local shrines are carved with mythological figures.

Temple ceremonials, especially the cremation of distinguished people, evoked elaborate ritual art objects in precious metals, as well as in wood or fabric. All were characterized by exuberant and repetitive curvilinear floral ornament and by figures based on Indian legend, especially the *Rāmāyaṇa* and parts of the *Mahābhārata*. In the villages today, music, dance, sculpture, and painting are focussed on the shrines and are practiced with an intensity unknown elsewhere in the world. Art is woven intimately into the life of the people. The masks carved of wood for the dances are specially refined, sometimes ornate versions of the masks used in the animist rituals of other Southeast Asian peoples. In the 20th century there are numerous village sculptors and painters, who sell to tourists work based upon the old ceremonial arts. During the 1930s an outside impetus to develop their traditional legendary imagery in Western formats came from a German painter, Walter Spies, who lived on Bali. A landscape tradition was evolved, and the painters have been able to communicate something of the extravagant visual charm of their island, giving glimpses of luminous village and mountain landscape. The style of both sculptors and painters, however, is based upon gently undulating curves and is often highly ornamental, with repeated patterns. A repertoire of posture and gesture has been abstracted from the *wayang*. The work thus tends to prettiness rather than vigour; the sculptors create no truly intelligible volumes, and the painters fill their surfaces with naïvely structured shapes.

Java: 20th century. A conscious revival of traditional art has been attempted in the 20th century, especially in Java, the main territory of modern Indonesia. There has been government support for the resuscitation of old crafts—silverwork, for example. A number of artists have adapted Westernized figure drawing to their own decorative compositions. The best known painter of Indonesia is the Javanese Affandi. He has used oil paint to execute pictures of Indonesian subjects in a vividly coloured Expressionist impasto (thick application of pigment to the canvas). This European brushwork technique, however, contains a strong element of the sinuosity of Javanese tradition. As yet, Affandi is the only artist from Southeast Asia to have attained a personal worldwide reputation.

Indonesian painting

THE PHILIPPINES

The population of this island group contains a number of different ethnic strata, the oldest of which shares in the general folk culture and its associated folk arts of the islands of Southeast Asia (see above *Indonesia*), with an emphasis on geometric simplification. An element in the Tagalog (a people of central Luzon) is perhaps descended from the oldest level of immigrants with a Paleolithic background. The Moro are Muslims, converted to Islām during the 15th and 16th centuries. Today they produce a decorative art in which old Muslim geometric motifs are combined with strong Chinese decorative influences (from Sung times, Chinese ceramics and textiles were imported). The decoration is applied primarily to textiles, weapons, and containers to hold the betel nuts that are chewed throughout Southeast Asia.

The most important departure in Philippine art was the result of the Spanish conquest of 1571. Thereafter, the bishopric of Manila and all of Luzon became the focus for an elaborate development of Spanish colonial art, primarily devoted to the construction and decoration of Roman Catholic churches in the current flamboyant, highly ornate, and colourful colonial style. There is good colonial architecture in other islands, including Bohol and Cebu. A large quantity of religious sculpture of the canonical Christian subjects was imported from Mexico and from Spain itself. Sculptors and missionary painters also immigrated, and a powerful local school developed under the direct influence of the 17th-century Spanish artists Murillo and Alonso Cano. Local arts were encouraged in 1785 by the remission of taxes for religious artists. Because of the close colonial ties, the stylistic developments corresponded

Influence of Spain upon Philippine art

San Agustin church, Manila, the Philippines, 1599–1614.
Bruce Coleman

substantially with those elsewhere in the Spanish Empire, and European prints served as models for local artists. Of the major early churches for which this sculpture and painting was executed, only San Agustin (1599–1614), in Manila, still stands; it was designed by Fray Antonio de Herrera, son or nephew of the great Spanish architect Juan de Herrera. During the 19th century the neo-Gothic style was imported, mainly through the Philippine architect Felipe Roxas, who had travelled in Europe and England. San Sebastian in Manila is a notable example of this style. The Spaniard Hervas, Manila's municipal architect from 1887 to 1893, favoured neo-Byzantine forms; *e.g.,* Manila Cathedral (1878–79).

It was only in the later 19th century that any secular art flourished at all. Schools of fine art modelled on the European schools were set up between 1815 and 1820, and a number of painters began to work in versions of European academic styles, painting landscapes, portraits, and classical subjects. The best known among them are Juan Luna, Felix Resurrección Hidalgo, Antonio Malantic, and the genre painter Fabian de la Rosa. After the transfer of rule to the United States in 1898, industrialization began in earnest; the methods of the art schools were adapted, as in Europe, to the needs of modern commercial society.

Modern Philippine architects

In the 1930s a substantial modern experimental school of Philippine architects began the remodelling of the industrial environment in terms of 20th-century architectural and design conceptions. Prominent names are Pablo Antonio, Carlos Arguelles, and Cesar Concio. Beginning in the 1930s but especially after World War II, artists in Manila adopted the Abstract and Expressionist styles current in the United States. After the devastation wrought by that war, Manila and other cities and towns were rebuilt, virtually anew, in local variants of the international business style.

FOLK ARTS

The arts of many regions in Southeast Asia remained either untouched or only slightly influenced by the Indianized arts of other regions. Such influence is found especially in regions where the gold trade flourished. In Sarawak (Bonkisam), for example, the remains of buildings similar to late Tantric east Javanese *tjandis* have been discovered. Among a few people (*e.g.,* the Meo of highland Vietnam), vestiges of Indian erotic temple imagery have been adapted to local fertility ceremonies, and most of the religious ideas of the region show at least faint traces of Indian influence.

Save for the megaliths and Dong Son bronzes, most of the known folk-art objects are relatively recent, although their inspiration and types belong to traditions far older and geographically more far-reaching than the Indianized traditions.

The two main non-Indian art styles in the whole region have been provisionally named the "monumental" and the "ornamental–fanciful." They coexist virtually everywhere, though they probably represent two evolutionary phases. The principal manifestations of the monumental style are the megalithic monuments, although there is great variety among the megalithic customs of the many different populations in Sumatra, Laos, Indonesia, Borneo, and the Philippines. The influence of the ornamental–fanciful style, which is characterized especially by the scrolled spiral, insinuates itself even into many of the decorative arts, particularly in the curvilinear and flamboyant inflection given to ornamental motives in the major Indianizing styles.

The squatting ancestor figure

The link between the two styles is probably the ubiquitous squatting ancestor figure, cocked knees supporting elbows, carved in soft wood or woven in cane or fibre. These figures may be either male or female. Under special social circumstances in recent times, very large wooden versions of the figure have been used as substitutes for more conventional, standing megalithic ancestral monuments (Sumatra and Sabah). The custom is probably an old one. There can be little doubt, for example, that the Theravāda Buddhist images of Burma, Thailand, and Laos were accepted as special modifications of the ancestor image. The transition from revering numinous ancestor images whose identity had been forgotten to worshipping an Indianizing icon was easy for the tribal populations.

The complex significance of the original squatting ancestor figure enabled it to be used in a variety of contexts. It might have combined associations of the fetus, the fetal burial position, and female birth and intercourse positions, as well as a ceremonial posture assumed by the living. It came to be used primarily in wooden sculpture on all scales, but also in woven textiles (*e.g.,* Iban), to represent the continuing power informing human existence, both in the purely ancestral sense of family continuity and identity and in the sense of the fertility of the land. Its earliest recorded appearance may be on Chinese Yang-shao painted pottery (*c.* 2000 BC); but it appears in essentially the same form over a range of territory including Sumatra, Nias and Sunda islands, Java, Borneo, New Guinea, Taiwan, the Philippines, and out into northern Australia and Melanesia. It may be used purely as an ancestral image in a family shrine house or as a motif added to any one of a variety of implements to potentiate them; for example, large bowls (Sumatra), kris or sword handles (Java, Sumatra, Borneo), spoons (Timor, the Philippines), musical instruments (Borneo), and magicians' staves (Borneo).

The treatment of such figures may be invested with more or less of the characteristics of the ornamental–fanciful style, in those regions where this style prevails (*e.g.,* Batak, Dayak). There are also special versions of the squatting figure that seem to belong especially to important magical crafts, such as the Javanese kris handle, on which miniature carvings can give an extraordinarily monumental effect. Sumatran Dayak hereditary magical staves may be carved with a "tower" or "tree" of such ancestor figures. On Nias, for example, along with the squatting figure, a standing figure in the bent-knee posture common in Polynesia also appears as a variant. In the Philippines similar variants are sometimes interpreted as vestiges from a remote Indian mythology, adopted probably for the sake of their cultural prestige. In southern Borneo the figure appears carved in the full round and as a pattern for woven textiles; it often has a protruding tongue and, sometimes, antlers—a combined motif known in the Ch'ang-sha art of southern China (*c.* 300 BC). Antlers also appear on certain Sumatran knifehilt figures. A variety of designs, some of them "abstract," are based on this figure. Among the Jarai of Vietnam, for example, a pattern of lozenges represents an abstraction from a group of these figures. Especially in the textiles of Sumba and other Indonesian islands, similar patterns, often referred to as decorated triangles,

represent the same phenomenon. When, as in textiles, the anthropomorphic reference of the abstract pattern is lost, the male genitals may remain to assert the ancestor significance.

Cult of the skull

The association between the squatting figure and the widely practiced cult of the skull is manifested in the combined cult of ancestors, headhunting, and head worship. Among the Wa of Upper Burma, for example, the squatting figure in a lozenge abstraction decorates the chests in which the severed heads of enemies are stored. Virtually everywhere among the early farmers of Southeast Asia, such heads were regarded as repositories of great spiritual power. The cult of the skull has produced a version of the squatting figure that is commonly known by the Indonesian word *korvar;* it is a figure with an ancestral skull in place of a carved head. Such figures are especially common in the more easterly island cultures. The ghostly power of the deceased ancestor can thus become present and available to his descendants—to give oracular advice, for example. A related idea is incorporated in the masks used in a wide variety of rituals and dance-dramas throughout Southeast Asia; for example, among the Batak of Sumatra and the Dayak of Borneo, where especially fine examples are made. There can be little doubt that the same idea (blended with imagery from the imported Hindu epics) underlies the range of elaborate masks that were once used in the Javanese and now can be seen in the Balinese *wayang* dances. It is possible that the flamboyant flame skull protuberances and winglike flanges ornamenting the head in so much of the Buddhist art produced in Burma and Thailand reflect a persistent but submerged interest in the cult of the skull.

Another major motif is the snake, which (even in areas where direct Indianizing influence was not strong) is frequently combined with imagery derived from the cult of the powerful, magical Hindu *nāga;* often many-headed, this serpent is the patron and guardian of water and treasure, both material and spiritual. The snake motif has also been blended with images of the Chinese dragon, going back perhaps to Chinese Han ornamental designs. Outstanding examples are found on the elaborate relief-carved doors of Sumatran Batak houses; "flying" roof finials in many parts of Indonesia; and in much Borneo Dayak ornament, from tattoos to carved bamboos and bronze body ornaments. The snake is the magico-mythical creature that gives both its bodily shape (either straight or undulant) and its metaphysical power to the kris. Distributed from Malacca to Celebes, these swords (the earliest known dated 1342) reached their high point of artistic development in Java. A variety of other motifs originating on the mainland of Asia is found in many of the surviving folk arts of Indonesia. Among them are the "man in the embrace of an animal" (Dayak kris handles) and animals "stacked" one above the other (Timor, Indonesia).

The ornamental–fantastic style. The styles in which these variations on basic motifs are carried out vary principally according to the preponderance of the sinuous curves and spirals of the ornamental–fantastic style. This style serves as the basis for decoration and as a method of artistic phrasing. It may have made its way into Southeast Asia as late as the 1st millennium BC, being formally related to the spirals used in Chinese Neolithic, Shang, and Chou bronze art. (It should, perhaps, be mentioned that Chinese art, certainly until well into the Christian Era, was itself far more "tribal" than later Chinese tradition recognizes.) Probably connoting spirituality, the spiral imagery appears in Southeast Asian magical art at all levels, from the textiles of Java and the incised bamboo implements or carved doors of Dayak Borneo to the ornament on the costumes of sculptured dancers or deities at every major city site. Given a fiery upward inflection, it appears in the finials on major Indianized stone architecture and on the carved wooden gables of Burmese and Thai Buddhist halls. There is not always complete stylistic consistency within any one cultural group. For example, the fantastic snake-dragon creatures carved in deep relief on the housedoors of the Batak may be extravagantly sinuous, with many spirals, while their figure sculpture adheres to the sterner plastic idiom, virtually without any

Use of curves and spirals

Brass receptacle in the shape of the mythical serpent, *naga*. From Krui, Sumatra. In the Royal Tropical Institute Museum, Amsterdam. Height 5 cm.

By courtesy of the Royal Tropical Institute, Amsterdam

linear sinuosity. Among the Dayak of Borneo the fantastic style may be confined entirely to surface ornament. On Indonesian islands, ancestral figures may be relatively static and foursquare, while the decorative carving and textiles may display considerable linear fantasy. A special version of the ornamental–fantastic style characterizes the surviving Indianized arts of Bali and Java, intruding even into sculptural inventions derived from strongly three-dimensional medieval Indianizing patterns. Thus, the decoration on the *wayang* cutout leather puppets, with its somewhat stereotyped curlicues, has proliferated at the expense of the three-dimensional sense (see above *Indonesia*). Balinese *wayang* masks may be carved entirely out of curling surfaces and completed in paint with sinuous eyebrows and moustaches. In many parts of Southeast Asia, including Thailand, Vietnam, Burma, Sumatra, and Indonesia, designs originally based upon Indian flowering-scroll patterns can be found in architecture, textiles, theatre costumes, musical instruments, and wooden utensils, all efflorescing with extravagant curling ornament. It is unfortunately true that only in a few of its most serious manifestations does this kind of ornament display substantial artistic invention, with carefully varied, asymmetrical, complementary, and counterchanged curves. Usually, it degenerates into repetitive space filling, without variety or formal meaning.

Textiles. Perhaps the types of folk art best known in the West are the textiles, especially batik and *ikat*. Both names refer to techniques practiced by different groups of people, who must have learned it from each other. Essentially Javanese but known in other islands, batik may have resulted from the imitation with dyes of South Indian painted cloths, probably before 1700. The essence of the technique is that melted wax is poured from a small metal kettle onto areas of a plain cotton cloth, which is then dyed, only the unwaxed parts taking the colour. The process can be repeated with several different colours. The oldest basic colours are indigo and brown; red and yellow were used later. The possible patterns range from lozenges and circlets through a large repertoire of cursive animal and plant forms. The batik technique can produce sumptuous and complex designs that not even the most elaborate weaving techniques can duplicate. It was encouraged by the Muslim rulers as a major element of social expression in garments and hangings.

Batik and *ikat*

Ikat is known among the Batak, in Cambodia, and especially among the dispersed Dayak people. It, too, probably originated in India. The extraordinarily difficult *ikat* textiles (woven cotton and occasionally silk, especially in Cambodia) are made primarily for use in important ceremonials and were regarded by their makers as major works of art. Before being woven, the thread is tightly tied at carefully calculated points in the hank (coiled or looped bundle); this is then dyed, the tied parts not taking up dye. The process may be repeated for different colours. As a consequence of the predyeing, designs appear as the thread is woven. In most *ikat* only the warp (the series

of yarns extended lengthwise in the loom and crossed by the weft) is so treated; but in southern Sumatra a tie-dyed floating weft is added to the plain weft. Naturally, *ikat* designs tend to be static and more or less rectilinear. In the finest *ikat*, however, birds and animals, spirits and houses, and, in Cambodia, a vestigial iconography of royal Buddhism may be formalized into extremely beautiful banded compositions. (P.S.R.)

BIBLIOGRAPHY. REGINALD LE MAY, *The Culture of Southeast Asia* (1954), deals with the art and architecture of Southeast Asian peoples. GEORGE COEDES, *Les États hindouisés d'Indochine et d'Indonésie*, new ed. (1964; Eng. trans., *The Indianized States of Southeast Asia*, 1968) and *Les Peuples de la péninsule indochinoise* (1962; Eng. trans., *The Making of Southeast Asia*, 1966), are standard works.

Literature: (*Burma*): MAUNG HTIN AUNG, *A History of Burma* (1967), contains a general survey of Burmese literature. The same author's *Burmese Folk-Tales* (1948), *Burmese Law Tales* (1962), *Burmese Monk's Tales* (1966); and MAUNG MYINT THEIN, *Burmese Folk-Songs* (1970), provide a good survey of Burmese oral literature. Dramatic literature is treated in MAUNG HTIN AUNG, *Burmese Drama* (1937). (*Thailand*): P. SCHWEISGUTH, *Étude sur la littérature siamoise* (1951), is the standard work on Thai literature. The *Journal of the Siam Society* (various issues) contains studies on Thai literature. (*The Philippines*): LEONARD CASPER, "The Philippine Literature: The Unexplored Potential," *Asia*, 9:80–87 (1967); ALBERT RAVENHOLT, *The Philippines: A Young Republic on the Move* (1962). (*Vietnam*): NGUYEN NGOC BICH (ed.), "The Poetry of Viet Nam," *Asia*, 14:69–91 (1969). (*Malaysia*): RICHARD WINSTEDT, *The Malays: A Cultural History*, 6th ed. (1961); and OLIVER RICE and ABDULLAH MAJID (eds.), *Modern Malay Verse, 1946–61* (1963), are important works to consult. (*Indonesia*): T.G.T. PIGEAUD, *Java in the 14th Century*, 5 vol. (1960–63), is a special study of Javanese literature in the Majapahit period. See also A. TEEUW, *Modern Indonesian Literature* (1967). TAKDIR ALISJAHBANA, *Indonesia in the Modern World* (1961), contains good accounts of modern Indonesian literature against the political, cultural, social, and historical background.

Music: WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East, and Asia*, ch. 2 and 5 (1967), discusses the music of the Southeast Asian region; LAWRENCE PICKEN, "The Music of Far Eastern Asia and Other Countries," in *The New Oxford History of Music*, vol. 1, *Ancient and Oriental Music*, pp. 83–194 (1957), discusses intercultural and musical relationships between Far Eastern and Southeast Asian countries. See also D.R. WIDDESS and R.F. WOLPERT (eds.), *Music and Tradition* (1981), essays on Asian music. (*Burma*): U KHIN ZAW, "Burmese Music: A Preliminary Inquiry," *Journal of the Burma Research Society*, pp. 387–466 (December 1940), appends several pages of music notations to a discussion of music history and the structure of Burmese scales. (*Cambodia and Laos*): ALAIN DANIELOU, "La Musique du Cambodge et du Laos" (1957), is a brochure that discusses with the help of illustrations traditional musical instruments of both countries. (*Indonesia*): JAAP KUNST, *Music in Java*, 2 vol. (1949), an important work with a comprehensive bibliography including works of many Dutch scholars, treats matters regarding history, vocal and instrumental music, structure, notation, and tonal systems of the East-Central and the West-Javanese gamelan; MANTLE HOOD, *The Nuclear Theme As a Determinant of Patet in Javanese Music* (1954), modal structure of patet is analyzed according to basic elements in themes of several gamelan pieces of music; COLIN MCPHEE, *Music in Bali* (1966), gives detailed descriptions and specific musical examples of repertoire played in many gamelan ensembles; WALTER KAUDERN, "Musical Instruments in Celebes," in *Ethnological Studies in Celebes*, vol. 3 (1927), a detailed study with a long list of names of musical instruments, 130 figures, and 19 maps showing the geographical distribution of musical instruments; JAAP KUNST, *Music in Nias* (1939), describes both vocal and instrumental music, classifies musical instruments (Hornbostel-Sachs divisions), and illustrates their distribution in 7 maps; *Music in Flores* (1942), lists in a table the native names of 54 instruments of the five divisions of the archipelago; CHARLES S. MYERS, "A Study of Sarawak Music," *Sammelbände der Internationalen Musikgesellschaft*, 15:296–308 (1913–14), 13 gongs from different cultural groups and music of some instruments are analyzed with the help of musical examples; HENRY L. ROTH, *The Natives of Sarawak and British North Borneo*, 2 vol. (1896)—vol. 1, ch. 9 deals with feasts, festivals and dancing, while vol. 2, ch. 26 discusses music in general; EDWIN H. GOMES, *Seventeen Years Among the Sea Dyaks of Borneo* (1911), terms for songs and names of musical instruments that the author cites are still known in Sarawak; IVOR EVANS, *Among Primitive Peoples in Borneo*, ch. 14 (1922), devoted exclusively to a discussion of musical instruments, music, and dancing; JAAP KUNST, *Music in New Guinea* (Eng. trans. 1967), comprises three works first published in 1931 and 1950 that treat of vocal and instrumental music of the Papua in the north and the central range of mountains, of songs in the north and the West, and of music in the West Central range, Southwest, North and West coasts (many musical examples and a distribution map of musical instruments). (*The Philippines*): JOSE MACEDA, "The Music of the Magindanao in the Philippines," (1963, Ann Arbor University Microfilms), discusses instrumental and vocal music of the whole culture with copious musical examples and two long-playing records available at Folkways Records, New York; and "Drone and Melody in Philippine Musical Instruments," paper read at an International Conference on Traditional Drama and Music of Southeast Asia, Kuala Lumpur, August 27–31 (1969), discusses age, spread, and variety of combinations of drone and melody as a root structure. (*Thailand*): DAVID MORTON, "The Traditional Instrumental Music of Thailand," (1964, Ann Arbor University Microfilms), a study that covers both the historical and structural aspects of Thai music. (*Vietnam*): TRAN-VAN-KHE, *La Musique Vietnamienne traditionelle* (1962), detailed and important information about history, musical instruments, and musical theory (references abound with criticism of some works); and *Vietnam* (1967), a clear and concise presentation, containing valuable new historical material.

Dance and theatre: JAMES R. BRANDON, *Theatre in Southeast Asia* (1967), a general survey of major theatre forms, incorporating firsthand observation. (*Burma*): MAUNG HTIN AUNG, *Burmese Drama* (1937), a detailed history and translations of plays, four complete and eight in excerpts; KENNETH SEIN and J.A. WITHEY, *The Great Po Sein* (1965), a lively chronicle of the Burmese stage of the last century, cast in the first person narrative, by a famous contemporary actor; U POK NI, *Konmara Pya Zat* (1952), a complete English translation of a 19th-century play with interpretive notes. (*Cambodia*): JACQUES BRUNET, "Nang Sbek, Danced Shadow Theatre of Cambodia," *World of Music*, 11:18–37 (1969), a brief and excellent description and history of Cambodian shadow theatre; SAMDACH CHAUFEA THIOUNN, *Danses cambodgiennes*, 2nd ed. (1956), the most complete history and analysis of plays, music, and dance of female and male dance-drama. (*Indonesia*): BENEDICT R.O'G. ANDERSON, *Mythology and the Tolerance of the Javanese* (1965), major *wajang* characters are described and their significance as behavioral models is treated; JAMES R. BRANDON (ed.), *On Thrones of Gold: Three Javanese Shadow Plays* (1970), translations, with description of action, music indication, and photographs, of three *wajang kulit* plays; CLAIRE HOLT, *Art in Indonesia* (1967), excellent chapters on *wajang kulit*, dance, and dance-drama in all parts of Indonesia; JAMES L. PEACOCK, *Rites of Modernization* (1968), the plays, structure and content, of *ludruk* in Java as seen through the eyes of a modern anthropologist; W.H. RASSERS, *Pañji, the Culture Hero* (1959), a disputed but brilliant theoretical discussion of the origin and meaning of Indonesian theatre; H. ULBRICHT, *Wayang Purwa: Shadows of the Past* (1970), useful for its extended translations of synopses of Pandawa plays found in J. KATS's famous Dutch work *Het Javaansche tooneel*, vol. 1, *Wajang Poerwa* (1923); BERYL DE ZOETE and WALTER SPIES, *Dance and Drama in Bali* (1938), an authoritative and encyclopaedic pre-World War II description of Bali's performing arts. See also I MADÉ BANDEM and FREDERIK E. DEBOER, *Kaja and Kelod: Balinese Dance in Transition* (1982); and ANA DANIEL, *Bali: Behind the Mask* (1981). (*Malaysia*): JEANNE CUISINIER, *Le Théâtre d'Ombres à Kelantan*, 2nd ed. (1957), Malaysian shadow theatre described and illustrated, with a partial play translation; RICHARD WINSTEDT, *The Malays: A Cultural History*, 6th ed. (1961), contains background information on Malaysian drama, including translations of pre-performance invocations. (*The Philippines*): JEAN EDADES (ed.), *Short Plays of the Philippines* (1950), a collection of recent one-act plays in English; ALBERTO S. FLORENTINO, *Outstanding Filipino Short Plays* (1961), short plays with two appendixes on traditional and modern drama in the Philippines; FRANCISCA TOLENTINO, *Philippine National Dances* (1946), descriptions and brief histories of many folk dances. (*Thailand*): UBOL BHUKKANASUT (trans.), "Manohra," in *Traditional Asian Plays*, ed. by JAMES R. BRANDON (1972), a translation with stage directions of the *lakon jatri* play "Manohra"; H.H. BRIDHYAKORN, "The Nang," 3rd ed. (1956), a short booklet on *nang yai* shadow play; and with DHANIT YUPHO, "The Khon," 3rd ed, (1956), a short booklet on *khon* masked pantomime; DHANIT YUPHO, *Classical Siamese Theatre* (1952), 52 folk and classical dance sequences described and illustrated; and *The Khōn and Lakon* (1963), synopses, commentary, and illustrations for 32 classic dance plays as performed by the Bangkok Department of Fine Arts between 1947 and 1960. (*Vietnam*): SONG-BAN, *The Vietnamese Theatre* (1960), brief descriptions of theatre in Vietnam.

Visual arts: PHILIP S. RAWSON, *The Art of Southeast Asia: Cambodia, Vietnam, Thailand, Laos, Burma, Java, Bali* (1967),

a comprehensive survey with many illustrations and plans; BENJAMIN ROWLAND, *The Art and Architecture of India,* 3rd ed. rev. (1967), sets the art of the region in relation to South Asian, or Indian art; WIM SWAAN, *Lost Cities of Asia* (1966), a pictorial study of Ceylon, Pagan, and Angkor; detailed articles in the *Encyclopedia of World Art* (1960–67): GEORGE COEDÈS, "Burmese Art," "Khmer Art," and "Cham Art"; A.B. GRISWOLD, "Siamese Art"; MADELEINE HALLADE and ROBERT HEINE-GELDERN, "Indonesian Art"; LOUIS BEZACIER, "Vietnamese Art"; and J.M.R. RIVIERE, "Philippine Art"—all these articles have extensive bibliographies. (*Burma*): A.B. GRISWOLD, CHEWON KIM, and P.H. POTT, *Burma, Korea, Tibet* (U.S. title, *The Art of Burma, Korea, Tibet;* 1964), the only survey of Burmese art; BURMA RESEARCH SOCIETY, *Fiftieth Anniversary Publications,* vol. 2 (1960), source material including important articles in English by G.H. LUCE and W.B. SINCLAIR. (*Champa*): The basic studies are PHILIPPE STERN, *L'Art du Champa (ancien Annam) et son évolution* (1942); and LOUIS BEZACIER, *Relevé des monuments anciens du Nord Viêt-nam* (1959). (*Siam and Laos*): Three basic sources include SILPA BHIRASRI, *Thai-Mon Bronzes* (1957), and *The Origin and Evolution of Thai Murals* (1959); and L. BORIBAL BURIBHAND and A.B. GRISWOLD, "Sculpture of Peninsular Siam in the Ayuthya Period," *Journal of the Siam Society,* 38:1–60 (1951). PIERRE DUPONT, *L'Archéologie mône de Dvāravatī* (1959), is the most authoritative study of this topic so far. The chief illustrated source for dating Buddhist art is A.B. GRISWOLD, *Dated Buddha Images of Northern Siam* (1957). See also CAROL STRATTON and MIRIAM M. SCOTT, *The Art of Sukhothai: Thailand's Golden Age* (1981), covering the period between the mid-1200s and the mid-1600s. (*Indochina*): JEAN BOISSELIER, *La Statuaire khmère et son évolution,* 2 vol. (1955), an exhaustive study of the development of Khmer sculpture; GEORGE COEDÈS, "Le Culte de la Royauté divinisée . . . ," *Série Orientale,* conference vol. 5 (1952), a basic iconographic study; PIERRE DUPONT, *La Statuaire préangkorienne* (1955), the authoritative book on pre-Angkor sculpture; LOUIS FREDERIC, *Sud-Est Asiatique: ses temples, ses sculptures* (1964; Eng. trans., *The Temples and Sculptures of Southeast Asia* (U.S. title, *The Art of Southeast Asia: Temples and Sculpture;* 1965), a well-documented pictorial survey; BERNARD P. GROSLIER, *Indochine: carrefour des arts* (1961; Eng. trans., *Indochina: Art in the Melting Pot of Races,* 1962), the most comprehensive survey in English; and with JACQUES ARTHAUD, *Angkor, hommes et pierres* (1956; Eng. trans., *Angkor: Art and Civilization,* rev. ed., 1966), a thoroughly documented pictorial survey. (*Indonesia*): F.A. WAGNER, *Indonesia: The Art of an Island Group* (1959), a comprehensive survey of Indonesian art in English; A.J. BERNET KEMPERS, *Ancient Indonesian Art* (1959), a comprehensive illustrated book. Major works on Borobuḍur include T. VAN ERP, *Beschrijving van Barabudur,* vol. 2, *Bouwkundige beschrijving* (1931); N.J. KROM, *Barabudur: Archaeological Description,* 2 vol. (1927); and PAUL MUS, *Barabudur, esquisse d'une histoire du bouddhisme fondée sur la critique archéologique des textes* (1935), which covers the subject of the role of Indian philosophy and theology as a background to Borobuḍur.

Southern Africa

The subcontinent of southern Africa is a vast region that includes the countries of Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Swaziland, Zambia, and Zimbabwe. (The island nation of Madagascar is excluded because of its distinct language and cultural heritage.) Southern Africa has been the background for a continuous political and economic struggle of European colonizers and the white elite of emerging nations with the diverse national groups seeking self-determination. Nowhere was this more in evidence than in South Africa, which since 1945 has been a frequent focus of world attention. The apartheid ("apartness") policy of the South African government, which established a system of strict racial segregation, evoked growing opposition from most other nations, especially as other prominent states in the area, such as Angola, Mozambique, and Zimbabwe (formerly Rhodesia), gained independence under majority rule. During this period, and in the 1980s in particular, the other southern African countries explored the possibilities of regional trade and cooperation as a means to their advancement. In the early 1990s South Africa dismantled its apartheid system and in 1994 held nonracial, democratic elections, thus opening the door to increased cooperation among all the nations of the region.

This article discusses the physical geography, settlement patterns, anthropology, economy, and history of the region of southern Africa, followed by separate treatment of the physical and human geography and history of each of the constituent states.

The article is divided into the following sections:

The region 838
  Physical and human geography 838
    The land
    The people
    The economy
  History 852
    Early humans and the Stone Age
    Food production
    The rise of more complex states
    The coming of the Portuguese
    The Dutch at the Cape
    The slave and ivory trade
    The "time of turmoil"
    British occupation of the Cape
    The expansion of white settlement
    Minerals and the scramble for southern Africa
    The South African War
    Southern Africa, 1910–45
    Southern Africa since 1945
The countries of southern Africa 874
  Angola 874
    Physical and human geography
    History
  Botswana 880
    Physical and human geography
    History
  Lesotho 886
    Physical and human geography
    History
  Malawi 890
    Physical and human geography
    History
  Mozambique 894
    Physical and human geography
    History
  Namibia 902
    Physical and human geography
    History
  South Africa 908
    Physical and human geography
    History
  Swaziland 929
    Physical and human geography
    History
  Zambia 933
    Physical and human geography
    History
  Zimbabwe 943
    Physical and human geography
    History
Bibliography 948

THE REGION

Physical and human geography

THE LAND

Geology. During the Triassic Period (245 to 208 million years ago) the continents formed a single landmass known as Pangaea. A southern continent, Gondwana (or Gondwanaland), formed about 180 million years ago as Pangaea began to divide. At this time Africa was attached to South America, but Antarctica had begun to split away. The separation of South America, Africa, and Madagascar began some 115 million years later.

The Sand River Gneisses of northern Transvaal formed about 3.8 billion years ago in the earliest part of the Precambrian (3.8 billion? to 540 million years ago) and are among the oldest rocks in the world. Other Precambrian rocks, as well as Cambrian and Ordovician rocks, are exposed over about a third of southern Africa. They are usually referred to as the basement complex, and, in most areas, the younger sedimentary rocks that mantle the basement complex have been metamorphosed or intruded by granite. The basement complex is well exposed on the Benguela and Niassa plateaus, in the Zambian-Zimbabwean Uplands, on the Damaraland Plain and southern African Lowveld, and in the Zambia-Niassa Trough and Limpopo Depression. Younger volcanic rocks that have intruded the Precambrian rocks occur throughout the basement complex. Important volcanic intrusions are the Witwatersrand System and Bushveld Igneous Complex in Transvaal and the Great Dyke in Zimbabwe. Linked to such intrusions are the economically important ore bodies and diamond pipes. Much of the basement complex has been uplifted and folded during periods of mountain building. The mountains formed during the Precambrian were eroded, and the resulting sediments were deposited on ancient erosion surfaces.

The basement complex

The rocks of the Karoo (Karroo) System were deposited between the Carboniferous (360 to 286 million years ago) and Early Jurassic (208 to 187 million years ago). They are mainly sediments derived from the erosion of basement-complex rocks under a variety of climates ranging from glacial to arid, although basaltic lavas (*e.g.*, Bataka basalts) are locally important. Later erosion and deposition have restricted outcroppings of the Karoo System to the rift valleys of Zambia and Zimbabwe and to an area encompassing much of southern Namibia and Cape Province and the Orange Free State in South Africa.

The Atlantic Coastal Plain in Namibia and Angola is bounded to the east by an ancient fault line, the Lunda Axis. There the granitic basement complex is overlain by folded and faulted rocks of the Cretaceous (144 to 66.4

million years ago) and Tertiary (66.4 to 1.6 million years ago), which in turn have been partially covered by Pliocene (5.3- to 1.6-million-year-old) sediments. The Mozambique Plain along the Indian Ocean coast probably formed during the detachment of Madagascar from the mainland, after which the plain was uplifted. There Cretaceous and Tertiary limestones, marls, and calcareous sandstones are overlain by Pliocene and Pleistocene (1,600,000- to 10,000-year-old) marine sands and clays.

The Kalahari Basin covers Botswana, northeastern Namibia, eastern Angola, and western Zambia. Its Tertiary and Quaternary rocks (those that formed from 1.6 million years ago to the present), which constitute the Kalahari System, have been deposited on Cretaceous sediments, which themselves were deposited on a Gondwana surface. These are for the most part sedimentary rocks derived from the erosion of the surrounding mountains. Fluvial (deposited by water) in origin, almost all the sediments have been reworked by the wind. The oldest rocks, the Botletle Sandstones and Grits, are overlain by the Kalahari Limestone Group, and overlying these are the Kalahari Sands. Sequences such as these can reach 1,000 feet (300 metres) in thickness.

Relief. The interior of southern Africa is dominated by a series of undulating plateaus that extend from the Cape provinces to central Angola. This area includes the southern African plateau, which covers most of South Africa, Namibia, and Botswana. Contiguous with this are the Zambian-Zimbabwean Uplands and Niassa and Manica plateaus in the northeast. This area is sometimes known as High Africa. The coastal margins are characterized by mountain ranges and coastal plains. Coastal mountains and escarpments, flanking the high ground, are well developed in northern Mozambique, South Africa, Namibia, Angola, and along the Mozambique-Zimbabwe border. Coastal plains abut the Indian Ocean in Mozambique and the Atlantic in Angola and Namibia.

Interior plateaus and coastal margins

The Kalahari Basin forms the central depression of the southern African plateau. Its elevation increases toward the plateau edge, and the highest ground, some 30 to 150 miles (50 to 240 kilometres) inland, forms the Great Escarpment that flanks the plateau in an almost unbroken line from the Zambezi River to Angola. The Great Escarpment was formed by prolonged uplift and includes ranges such as the Drakensberg and Stormberg and the Sneeu and Nuweveld mountains. The relative relief of some mountain ranges at the plateau edge is accentuated by the occurrence of deposits of Mesozoic lava (those formed from about 245 to 66.4 million years ago). The Drakensberg, which include the region's highest mountain—Lesotho's Mount Ntlenyana at 11,424 feet (3,482 metres)—are partially formed from such lava.

The plateau is divided into three altitudinal zones: the Highveld at altitudes greater than 4,000 feet, the Middleveld from 2,000 to 4,000 feet, and the Lowveld below 2,000 feet. The Highveld is well developed in the southern Transvaal region and Orange Free State, with important outliers such as the dolomitic Kaap Plateau, the Basotho Highlands, and the Transvaal Bushveld Basin, which is dominated by the Bushveld Igneous Complex. The Middleveld merges with the Highveld in most areas and is best developed in the western part of the region. It extends from the Namaqualand Plain northward through Namibia and into central Angola as far as the Bié Plateau. It includes the Namaland, Khomas, Otavi, and Chela highlands in Namibia, all of which are developed on the basement complex. The maximum extent of the Lowveld is in the Kalahari desert, the largest continuous sand surface in the world. Many dune fields are no longer active, and the fields of long parallel dunes have been attributed to drier paleoclimates that prevailed in the region in the late Quaternary. Saline depressions—pans—that occasionally fill with water are common landforms, although they vary greatly in size. The area includes two of the largest salt flats in the world—the Makgadikgadi and Etosha pans in northern Botswana and northern Namibia, respectively. The occurrence of river and lake sediments from these areas also attests to periods of wetter paleoclimates.

The Zambian-Zimbabwean Uplands are contiguous with the southern African plateau and form a flat planation surface at an altitude of 4,000 to 5,500 feet. The monotony is occasionally broken by isolated hills (inselbergs), mountain ranges (*e.g.,* the Muchinga Mountains in northern Zambia), and volcanic intrusions (such as Zimbabwe's long Great Dyke). On the plateaus most valleys are shallow, and the heads of drainage networks often form wide, streamless depressions known as dambos. Shallow lakes and swamps, such as Lake Bangweulu and the Kafue Flats in Zambia, are common. The only substantial valleys are formed by the Zambezi and Luangwa (Aruângua) rivers. Their courses are often fault-controlled, which has led to the creation of deep valleys, sometimes as much as 3,300 feet in depth. Dissection and local lithology often combine to create spectacles such as the Victoria Falls. To the east of the Zambian-Zimbabwean Uplands are the related Niassa and Manica plateaus, which form an undulating surface between 1,300 and 4,000 feet in altitude. These plateaus are formed from the basement complex, and

Brian Seed/Tony Stone Worldwide

Cattle watering on the rolling farmland of the southern African plateau in Eastern Cape province, South Africa.

MAP INDEX

Cities and towns

Alexander Bay . . . . 28 39 s 16 31 E
Alice . . . . . 32 47 s 26 50 E
Aliwal North . . . . . 30 42 s 26 42 E
Ancuabe . . . . . 12 58 s 39 51 E
Angoche . . . . . 16 15 s 39 54 E
Arandis . . . . . 22 25 s 14 58 E
Aranos . . . . . 24 08 s 19 07 E
Bagani . . . . . 18 07 s 21 38 E
Balaka . . . . . 14 59 s 34 57 E
Bancroft, see Chililabombwe
Banket . . . . . 17 23 s 30 24 E
Beaufort West . . . . 32 21 s 22 35 E
Beira . . . . . 19 50 s 34 52 E
Beitbridge . . . . . 22 13 s 30 00 E
Bellville . . . . . 33 54 s 18 38 E
Benguela (São Félipe de Benguela) . . . . . 12 35 s 13 24 E
Bethanien . . . . . 26 29 s 17 09 E
Bethlehem . . . . . 28 14 s 28 18 E
Bindura . . . . . 17 18 s 31 20 E
Bisho . . . . . 32 53 s 27 24 E
Blantyre . . . . . 15 47 s 35 00 E
Bloemfontein . . . . . 29 08 s 26 10 E
Bobonong . . . . . 21 58 s 28 20 E
Boksburg . . . . . 26 13 s 28 15 E
Bredasdorp . . . . . 34 32 s 20 02 E
Bremersdorp, see Manzini
Broken Hill, see Kabwe
Buffalo Range . . . . 21 01 s 31 32 E
Bulawayo . . . . . 20 09 s 28 35 E
Butterworth . . . . . 32 20 s 28 09 E
Caála (Robert Williams) . . . . . 12 51 s 15 34 E
Cabinda . . . . . 05 33 s 12 12 E
Cacolo . . . . . 10 08 s 19 16 E
Caconda . . . . . 13 44 s 15 04 E
Caluquembe . . . . . 13 52 s 14 26 E
Calvinia . . . . . 31 28 s 19 47 E
Camacupa (General Machado) . . . . . 12 01 s 17 29 E
Cangamba . . . . . 13 41 s 19 52 E
Cape Town (Kaapstad) . . . . . 33 55 s 18 25 E
Carletonville . . . . . 26 22 s 27 24 E
Carmona, see Uíge
Carnarvon . . . . . 30 57 s 22 08 E
Catandica . . . . . 19 09 s 33 12 E
Catumbela . . . . . 12 26 s 13 33 E
Chakari . . . . . 18 04 s 29 51 E
Chegutu (Hartley) . . 18 08 s 30 09 E
Chibuto . . . . . 24 42 s 33 33 E
Chikwawa . . . . . 16 03 s 34 48 E
Chileka . . . . . 14 01 s 33 24 E
Chililabombwe (Bancroft) . . . . . 12 22 s 27 50 E
Chimanimani (Mandidzudzure, or Melsetter) . . . . . 19 48 s 32 52 E
Chimoio (Vila Pery) . . . . . 19 08 s 33 29 E
Chinde, see Vila do Chinde
Chingola . . . . . 12 32 s 27 52 E
Chinhoyi (Sinoia) . . 17 22 s 30 12 E
Chipata (Fort Jameson) . . . 13 39 s 32 40 E
Chipinge . . . . . 20 12 s 32 37 E
Chiredzi . . . . . 21 03 s 31 40 E
Chitungwiza . . . . . 18 47 s 32 37 E
Chiúre . . . . . 13 22 s 39 58 E
Chokwe . . . . . 24 32 s 32 59 E
Cholo (Thyolo) . . . . 16 04 s 35 08 E
Choma . . . . . 16 49 s 26 59 E
Cuamba, see Nova Freixo
Cubal . . . . . 13 02 s 14 15 E
Cuchi . . . . . 14 39 s 16 54 E
Dalatando, see N'dalatando
Damba . . . . . 06 41 s 15 08 E
De Aar . . . . . 30 39 s 24 01 E
Dedza . . . . . 14 22 s 34 20 E
Dowa . . . . . 13 39 s 33 56 E
Dundee . . . . . 28 10 s 30 14 E
Dundo . . . . . 09 48 s 14 41 E
Durban (Port Natal) . . . . . 29 51 s 31 01 E
East London . . . . . 33 02 s 27 55 E
Empangeni . . . . . 28 45 s 31 54 E
Empress Mine Township . . . . . 18 27 s 29 27 E
Entre Rios (Malêma) . . . . . 14 57 s 37 25 E
Ermelo . . . . . 26 32 s 29 59 E
Fingoè . . . . . 15 10 s 31 53 E
Fort Jameson, see Chipata
Fort Johnson, see Mangoche
Fort Manning, see Mchinji
Fort Rixon . . . . . 20 01 s 29 16 E
Fort Rosebery, see Mansa
Fort Victoria, see Masvingo
Francistown . . . . . 21 13 s 27 31 E
Gabela . . . . . 10 51 s 14 22 E
Gaborone . . . . . 24 40 s 25 54 E
Ganda (Mariano Machado) . . . . . 13 01 s 14 38 E
Gatooma, see Kadoma
General Machado, see Camacupa
George . . . . . 33 58 s 22 27 E
Germiston . . . . . 26 13 s 28 11 E
Ghanzi . . . . . 21 34 s 21 47 E
Giyani . . . . . 23 19 s 30 43 E
Glendale . . . . . 17 21 s 31 04 E
Gobabis . . . . . 22 27 s 18 58 E
Gochas . . . . . 24 47 s 18 49 E
Graaff-Reinet . . . . . 32 15 s 24 33 E
Grahamstown . . . . 33 18 s 26 32 E
Grootfontein . . . . . 19 34 s 18 07 E
Gurué, see Vila Junqueiro
Gweru (Gwelo) . . . . 19 27 s 29 49 E
Hammarsdale . . . . . 29 48 s 30 39 E
Harare (Salisbury) . . 17 50 s 31 03 E
Hartley, see Chegutu
Henrique de Carvalho, see Saurimo
Hlatikulu . . . . . 26 58 s 31 19 E
Hopefield . . . . . 33 04 s 18 21 E
Hopetown . . . . . 29 37 s 24 05 E
Huambo (Nova Lisboa) . . . . . 12 46 s 15 44 E
Humansdorp . . . . . 34 02 s 24 46 E
Hwange (Wankie) . . 18 22 s 26 29 E
Inhambane . . . . . 23 52 s 35 23 E
Inyanga . . . . . 18 13 s 32 45 E
Isoka . . . . . 10 08 s 32 38 E
Jagersfontein . . . . . 29 46 s 25 25 E
João Belo, see Xai Xai
Johannesburg . . . . 26 12 s 28 05 E
Jwaneng . . . . . 24 35 s 24 42 E
Kaapstad, see Cape Town
Kabwe (Broken Hill) . . . . . 14 27 s 28 27 E
Kadake Station . . . . 26 13 s 31 02 E
Kadoma (Gatooma) . . . . . 18 21 s 29 55 E
Kafue . . . . . 15 47 s 28 11 E
Kalabo . . . . . 14 58 s 22 41 E
Kalamare . . . . . 22 52 s 26 30 E
Kalkfeld . . . . . 20 53 s 16 11 E
Kalulushi . . . . . 12 50 s 28 05 E
Kanye . . . . . 24 59 s 25 21 E
Kapiri Mposhi . . . . . 13 58 s 28 41 E
Karasburg . . . . . 28 01 s 18 45 E
Kariba . . . . . 16 31 s 28 48 E
Karibib . . . . . 21 56 s 15 50 E
Karoi . . . . . 16 49 s 29 41 E
Karonga . . . . . 09 56 s 33 56 E
Kasama . . . . . 10 13 s 31 12 E
Kasane . . . . . 17 49 s 25 09 E
Kasungu . . . . . 13 02 s 33 29 E
Katete . . . . . 14 06 s 32 05 E
Kathu (Sishen) . . . . 27 40 s 23 01 E
Kawambwa . . . . . 09 47 s 29 05 E
Keetmanshoop . . . . 26 35 s 18 08 E

DEMOCRATIC REPUBLIC OF THE CONGO
TANZANIA
ZAMBIA
MOZAMBIQUE
ZIMBABWE
BOTSWANA
SOUTH AFRICA
SWAZILAND
MADAGASCAR
KENYA
BURUNDI
WESTERN
RIFT
VALLEY
MOZAMBIQUE PLATEAU
MOZAMBIQUE PLAIN
HIGHVELD
LOWVELD
Mozambique Channel
INDIAN OCEAN
Bujumbura
Mombasa
Tabora
Tanga
Zanzibar
Dar es Salaam
Kananga
Mbuji-Mayi
Lubumbashi
Kasama
Mansa
Chililabombwe
Chingola
Kalulushi
Luanshya
Mufulira
Kitwe
Ndola
Kabwe
Lusaka
Kafue
Mongu
Senanga
Livingstone
Chipata
Mzuzu
Lilongwe
Lichinga
Zomba
Blantyre
Tete
Pemba
Nacala
Nampula
Mahajanga
Quelimane
Harare
Chitungwiza
Chegutu
Marondera
Kadoma
Kwekwe
Gweru
Mutare
Masvingo
Bulawayo
Hwange
Chimoio
Beira
Francistown
Selebi-Phikwe
Gaborone
Pietersburg
Rustenburg
Pretoria
Johannesburg
Soweto
Germiston
Boksburg
Roodepoort
Krugersdorp
Vereeniging
Vanderbijlpark
Welkom
Kimberley
Bloemfontein
Maseru
Pietermaritzburg
Durban
Pinetown
Newcastle
Maputo
Matola
Mbabane
Inhambane
Xai Xai
East London
Port Elizabeth
Oudtshoorn
George
Mongu
KAFUE NATIONAL PARK
CHOBE NATIONAL PARK
HWANGE NATIONAL PARK
CENTRAL KALAHARI GAME RESERVE
KRUGER NATIONAL PARK
GORONGOSA NATIONAL PARK
BANHINE NATIONAL PARK
Cities over 1,000,000
Cities 250,000 to 1,000,000
Cities 30,000 to 250,000
Cities under 30,000
Other localities
National capitals
International boundaries
Intermittent rivers
Reefs
Dams
Waterfalls
Salt lakes
Swamps and marshes
Flooded areas
Salt flats
Sand areas
Points of interest
Spot elevations in metres (1 m = 3.28 ft)
Scale 1 : 14,682,000
1 inch equals approx. 232 miles
0 50 150 250 350 450 km
Lambert Zenithal Equal Area Projection

Khonxas ..... 20 22 S 14 58 E
Kimberley ..... 28 45 S 24 46 E
King William's Town, see Bisho
Kitwe ..... 12 49 S 28 13 E
Klerksdorp ..... 26 52 S 26 40 E
Kota Kota, see Nkhota Kota
Kroonstad ..... 27 40 S 27 14 E
Krugersdorp ..... 26 06 S 27 46 E
Kuito (Silva Porto) ..... 12 23 S 16 56 E
Kuruman ..... 27 28 S 23 26 E
Kwekwe (Que Que) ..... 18 55 S 29 49 E
Ladysmith ..... 28 33 S 29 47 E
Lebowakgomo ..... 24 12 S 29 30 E
Leonardville ..... 23 30 S 18 48 E
Letlhakane ..... 21 25 S 25 35 E
Lichinga ..... 13 18 S 35 14 E
Lilongwe ..... 13 59 S 33 47 E
Livingstone (Maramba) ..... 17 51 S 25 52 E
Lobamba ..... 26 28 S 31 12 E
Lobatse ..... 25 13 S 25 40 E
Lobito ..... 12 21 S 13 33 E
Lourenço Marques, see Maputo
Luanda (São Paulo de Luanda) ..... 08 49 S 13 15 E
Luanshya ..... 13 08 S 28 25 E
Luau ..... 10 42 S 22 14 E
Lubango (Sá da Bandeira) ..... 14 55 S 13 30 E
Lucapa ..... 08 25 S 20 45 E
Lüderitz ..... 26 38 S 15 09 E
Luena (Vila Luso) ..... 11 47 S 19 55 E
Lusaka ..... 15 25 S 28 17 E
Mafeteng ..... 29 49 S 27 15 E
Mafikeng (Mafeking) ..... 25 52 S 25 39 E
Magude ..... 25 02 S 32 40 E
Mahalapye ..... 23 04 S 26 50 E
Malanje ..... 09 32 S 16 20 E
Malêma, see Entre Rios
Maltahöhe ..... 24 50 S 16 59 E
Mandidzudzure, see Chimanimani
Mangoche (Fort Johnson) ..... 14 28 S 35 16 E
Mansa (Fort Rosebery) ..... 11 12 S 28 53 E
Manzini (Bremersdorp) ..... 26 29 S 31 22 E
Maputo (Lourenço Marques) ..... 25 58 S 32 34 E
Maramba, see Livingstone
Marandellas, see Marondera
Margate ..... 30 51 S 30 22 E
Mariano Machado, see Ganda
Mariental ..... 24 38 S 17 58 E
Marondera (Marandellas) ..... 18 11 S 31 33 E
Marromeu ..... 18 17 S 35 56 E
Maseru ..... 29 19 S 27 29 E
Mashava ..... 20 03 S 30 29 E
Massinga ..... 23 20 S 35 22 E
Masvingo (Fort Victoria, or Nyanda) ..... 20 05 S 30 50 E
Matola ..... 25 58 S 32 28 E
Maun ..... 19 59 S 23 25 E
Mavinga ..... 15 48 S 20 21 E
Mazabuka ..... 15 51 S 27 46 E
Mbabane ..... 26 19 S 31 08 E
Mbala ..... 08 50 S 31 22 E
M'banza Congo (São Salvador) ..... 06 16 S 14 15 E
Mchinji (Fort Manning) ..... 13 48 S 32 54 E
Melsetter, see Chimanimani
Memba ..... 14 12 S 40 32 E
Menongue (Serpa Pinto) ..... 14 40 S 17 42 E
Messina ..... 22 21 S 30 03 E
Mhangura ..... 16 54 S 30 09 E
Middelburg ..... 25 47 S 29 28 E
Mlanje, see Mulanje
Mmabatho ..... 25 51 S 25 38 E
Moçambique (Mozambique) ..... 15 03 S 40 45 E
Moçâmedes, see Namibe
Mochudi ..... 24 25 S 26 09 E
Mocubúri ..... 14 39 S 38 54 E
Mohales Hoek ..... 30 09 S 27 28 E
Molepolole ..... 24 25 S 25 32 E
Moma ..... 16 44 S 39 14 E
Mongu ..... 15 17 S 23 08 E
Monkey Bay ..... 14 05 S 34 55 E
Monze ..... 16 16 S 27 29 E
Mopeia Velha ..... 17 59 S 35 43 E
Morrumbala ..... 17 20 S 35 35 E
Morrumbene ..... 23 39 S 35 20 E
Mossamedes, see Namibe
Mossel Bay ..... 34 11 S 22 08 E
Mount Darwin ..... 16 47 S 31 35 E
Mozambique, see Moçambique
Mpika ..... 11 50 S 31 27 E
Mpulungu ..... 08 46 S 31 07 E
Mufulira ..... 12 33 S 28 14 E
Mulanje (Mlanje) ..... 16 02 S 35 30 E
Mumbwa ..... 14 59 S 27 04 E
Murrupula ..... 15 27 S 38 47 E
Mutare (Umtali) ..... 18 58 S 32 40 E
Mvuma ..... 19 17 S 30 32 E
Mwamfuli (Samfya) ..... 11 21 S 29 33 E
Mzimba ..... 11 54 S 33 36 E
Mzuzu ..... 11 27 S 33 55 E
Nacala ..... 14 33 S 40 40 E
Namapa ..... 13 43 S 39 50 E
Namibe (Moçâmedes, or Mossamedes) ..... 15 10 S 12 09 E
Nampula ..... 15 09 S 39 18 E
Nchelenge ..... 09 21 S 29 44 E
N'dalatando (Dalatando, or Salazar) ..... 09 18 S 14 55 E
Ndola ..... 12 58 S 28 38 E
Negage ..... 07 46 S 15 16 E
Nelspruit ..... 25 28 S 30 58 E
Newcastle ..... 27 45 S 29 56 E
Nhlangano ..... 27 07 S 31 12 E
Nkhata Bay ..... 11 36 S 34 18 E
Nkhota Kota (Kota Kota) ..... 12 55 S 34 18 E
Nóqui ..... 5 51 S 13 26 E
Norton ..... 17 53 S 30 42 E
Nova Freixo (Cuamba) ..... 14 49 S 36 33 E
Nova Lisboa, see Huambo
Nova Mambone ..... 20 58 S 35 00 E
Novo Redondo, see Sumbe
Nsanje (Port Herald) ..... 16 55 S 35 16 E
Nyanda, see Masvingo
Odendaalsrus ..... 27 52 S 26 41 E
Okahandja ..... 21 59 S 16 55 E
Omaruru ..... 21 26 S 15 56 E
Ondangwa (Ondangua) ..... 17 55 S 15 57 E
Ondjiva ..... 17 04 S 15 44 E
Opuwo ..... 18 04 S 13 51 E
Oranjemund ..... 28 33 S 16 26 E
Orapa ..... 21 17 S 25 22 E
Orkney ..... 26 59 S 26 40 E
Oshakati ..... 17 47 S 15 41 E
Otavi ..... 19 39 S 17 20 E
Otjimbingwe ..... 22 21 S 16 08 E
Otjiwarongo ..... 20 27 S 16 39 E
Oudtshoorn ..... 33 35 S 22 12 E
Outjo ..... 20 07 S 16 09 E
Paarl ..... 33 44 S 18 58 E
Palapye (Palapye Road) ..... 22 33 S 27 08 E
Panda ..... 24 03 S 34 43 E
Pemba ..... 12 57 S 40 30 E
Petauke ..... 14 15 S 31 20 E
Phalaborwa ..... 23 57 S 31 07 E
Phuthaditjhaba, see Witsieshoek
Piet Retief ..... 27 00 S 30 48 E
Pietermaritzburg ..... 29 37 S 30 23 E
Pietersburg ..... 23 54 S 29 27 E
Piggs Peak ..... 25 58 S 31 15 E
Pilgrim's Rest ..... 24 53 S 30 45 E
Pinetown ..... 29 49 S 30 51 E
Plumtree ..... 20 29 S 27 49 E
Port Elizabeth ..... 33 58 S 25 35 E
Port Herald, see Nsanje
Port Natal, see Durban
Port Nolloth ..... 29 15 S 16 52 E
Port Shepstone ..... 30 45 S 30 27 E
Porto Alexandre, see Tombua
Porto Amboin ..... 10 44 S 13 45 E
Potchefstroom ..... 26 43 S 27 06 E
Pretoria ..... 25 45 S 28 10 E
Prieska ..... 29 40 S 22 45 E
Que Que, see Kwekwe
Queenstown ..... 31 54 S 26 53 E
Quelimane ..... 17 51 S 36 52 E
Quimbele ..... 06 31 S 16 13 E
Quissico ..... 24 43 S 34 45 E
Quthing ..... 30 24 S 27 43 E
Ramotswa ..... 24 52 S 25 49 E
Redcliff ..... 19 02 S 29 47 E
Rehoboth ..... 23 19 S 17 05 E
Ribauè ..... 14 57 S 38 19 E
Richard's Bay ..... 28 48 S 32 05 E
Robert Williams, see Caála
Roma ..... 29 27 S 27 42 E
Roodepoort ..... 26 10 S 27 52 E
Rundu ..... 17 56 S 19 46 E
Rusape ..... 18 44 S 32 02 E
Rustenburg ..... 25 40 S 27 15 E
Sá da Bandeira, see Lubango
St. Helena Bay ..... 32 46 S 18 02 E
Salazar, see N'dalatando
Salima ..... 13 47 S 34 26 E
Salisbury, see Harare
Samfya, see Mwamfuli
Santa Comba, see Waku Kungo
São Félipe de Benguela, see Benguela
São Paulo de Luanda, see Luanda
São Salvador, see M'banza Congo
Sasolburg ..... 26 49 S 27 49 E
Saurimo (Henrique de Carvalho) ..... 09 39 S 20 24 E
Selebi-Phikwe ..... 22 01 S 27 50 E
Selukwe (Shurugwi) ..... 19 40 S 30 00 E
Senanga ..... 16 07 S 23 16 E
Serenje ..... 13 14 S 30 14 E
Serowe ..... 22 23 S 26 43 E
Serpa Pinto, see Menongue
Seshego ..... 23 51 S 29 23 E
Shabani, see Zvishavane
Shamva ..... 17 19 S 31 34 E
Shashe ..... 21 26 S 27 27 E
Shurugwi, see Selukwe
Sibasa ..... 22 56 S 30 29 E
Silva Porto, see Kuito
Simonstown ..... 34 12 S 18 27 E
Sinoia, see Chinhoyi
Sishen, see Kathu
Siteki (Stegi) ..... 26 27 S 31 57 E
Solwezi ..... 12 11 S 26 24 E
Songo ..... 15 36 S 32 44 E
Soweto ..... 26 16 S 27 52 E
Soyo ..... 06 08 S 12 22 E
Stampriet ..... 24 20 S 18 24 E
Stegi, see Siteki
Stellenbosch ..... 33 56 S 18 51 E
Sumbe (Novo Redondo) ..... 11 12 S 13 50 E
Swakopmund ..... 22 41 S 14 32 E
Swellendam ..... 34 02 S 20 26 E
Tembisa ..... 25 59 S 28 13 E
Tete ..... 16 10 S 33 36 E
Teyateyaneng ..... 29 09 S 27 44 E
Thabazimbi ..... 24 36 S 27 24 E
Thohoyandou ..... 22 57 S 30 29 E
Thyolo, see Cholo
Tlokweng ..... 24 32 S 25 58 E
Tombua (Porto Alexandre) ..... 15 48 S 11 51 E
Tongaat ..... 29 35 S 31 08 E
Triangle ..... 21 02 S 31 27 E
Tshabong ..... 26 03 S 22 27 E
Tshane ..... 24 05 S 21 54 E
Tsumeb ..... 19 14 S 17 43 E
Tsumkwe ..... 19 36 S 20 30 E
Tulbagh ..... 33 17 S 19 09 E
Tuli ..... 21 55 S 29 12 E
Uíge (Carmona) ..... 07 37 S 15 03 E
Uitenhage ..... 33 46 S 25 24 E
Ulonguè ..... 14 43 S 34 21 E
Ulundi ..... 28 17 S 31 25 E
Umlazi ..... 29 49 S 29 59 E
Umtali, see Mutare
Umtata ..... 31 35 S 28 47 E
Upington ..... 28 27 S 21 15 E
Usakos ..... 22 00 S 15 36 E
Utrecht ..... 27 39 S 30 20 E
Vanderbijlpark ..... 26 42 S 27 49 E
Vereeniging ..... 26 40 S 27 56 E
Victoria Falls ..... 17 56 S 25 50 E
Vila da Manhiça ..... 25 24 S 32 48 E
Vila da Mocimboa da Praia ..... 11 20 S 40 21 E
Vila de Mocuba ..... 16 51 S 36 56 E
Vila do Chinde (Chinde) ..... 18 34 S 36 27 E
Vila do Zumbo ..... 15 36 S 30 25 E
Vila Junqueiro (Gurué) ..... 15 28 S 36 59 E
Vila Luso, see Luena
Vila Pery, see Chimoio
Vilankulo (Vilanculos) ..... 22 00 S 35 19 E
Virginia ..... 28 07 S 26 54 E
Vredenburg ..... 32 54 S 17 59 E
Vryburg ..... 26 57 S 24 44 E
Vryheid ..... 27 46 S 30 48 E
Waku Kungo (Santa Comba) ..... 11 21 S 15 07 E
Walvis Bay ..... 22 57 S 14 30 E
Wankie, see Hwange
Warmbad ..... 28 27 S 18 44 E
Welkom ..... 27 59 S 26 42 E
Wellington ..... 33 38 S 19 00 E
Windhoek ..... 22 35 S 17 05 E
Witbank ..... 25 52 S 29 14 E
Witsieshoek (Phuthaditjhaba) ..... 28 32 S 28 48 E
Worcester ..... 33 39 S 19 26 E
Xai Xai (Joaõ Belo) ..... 25 04 S 33 39 E
Zambezi ..... 13 33 S 23 07 E
Zomba ..... 15 23 S 35 20 E
Zvishavane (Shabani) ..... 20 20 S 30 02 E
Zwelitsha ..... 32 55 S 27 25 E

Physical features and points of interest

Agulhas, Cape ..... 34 50 S 20 00 E
Aha, Mount ..... 19 47 S 21 06 E
Algoa Bay ..... 33 50 S 25 50 E
Angolan, see Bié Plateau
Aruângua, see Luangwa
Atlantic Ocean ..... 28 00 S 12 00 E
Auas Mountains ..... 22 38 S 17 12 E
Augrabies Falls ..... 28 35 S 20 23 E
Bangweulu, Lake ..... 11 05 S 29 45 E
Bangweulu Swamps ..... 11 30 S 30 15 E
Banhine National Park ..... 22 45 S 33 00 E

Barotse Flood Plain . . . . . 15 40 S 23 10 E
Bazaruto Island . . . 21 40 S 35 25 E
Bicuari (Bikaur) National Park . . . . 15 20 S 14 50 E
Bié (Angolan) Plateau . . . . . 12 30 S 17 00 E
Binga, Mount . . . . . 19 20 S 32 47 E
Blyde, *river* . . . . . 24 16 S 30 50 E
Bokkeveld, Mount . 31 10 S 18 56 E
Boteti (Botletle), *river* . . . . . 20 08 S 23 33 E
Brand, Mount . . . . . 21 08 S 14 35 E
Brukkaros, Mount . . . . . 25 52 S 17 48 E
Bua, *river* . . . . . 12 45 S 34 16 E
Busanga Swamp . . 14 10 S 25 50 E
Búzi (Busi), *river* . . . . 19 52 S 34 46 E
Cahora Bassa, Lake . . . . . 15 34 S 32 50 E
Cahora Bassa Dam . . . . . 15 35 S 32 44 E
Caledon (Mohokare), *river* . . . . . 30 32 S 26 05 E
Caprivi Strip, *region* . . . . . 18 00 S 23 00 E
Catumbela (Katumbela), *river* . . . . . 12 27 S 13 29 E
Central Kalahari Game Reserve . . . 22 00 S 24 00 E
Central Karoo, see Great Karoo
Chambeshi, *river* . . 11 53 S 29 48 E
Changane, *river* . . . 24 43 S 33 32 E
Chela, Serra da . . . 15 30 S 13 30 E
Chewore Safari Park . . . . . 16 00 S 30 00 E
Chicapa, *river* . . . . 06 25 S 20 48 E
Chilengue, Serra . . . 13 18 S 15 22 E
Chilwa (Chirua, or Shirwa), Lake . . . . . 15 12 S 35 50 E
Chimoio Plateau . . . 19 00 S 33 30 E
Chire, see Shire
Chizarira Hills . . . . . 17 43 S 27 45 E
Chobe, see Linyanti
Chobe National Park . . . . . 18 30 S 24 30 E
Correntes Point . . . 24 07 S 35 30 E
Crocodile, see Krokodil
Cuando, see Kwando
Cuango, see Kwango
Cuanza, see Kwanza
Cubango, see Okavango
Cuito, *river* . . . . . 18 01 S 20 48 E
Cunene, see Kunene
Cuvo, *river* . . . . . 10 52 S 13 48 E
Dande, *river* . . . . . 08 28 S 13 21 E
Delagoa (Lourenço Marques) Bay . . . 25 48 S 32 51 E
Delgado, Cape . . . . 10 40 S 40 38 E
Dow, see Xau, Lake
Drakensberg, *mountains* . . . . 29 00 S 29 00 E
Dwangwa Estates, *point of interest* . . . . . 12 30 S 34 10 E
East African Rift Valley . . . 14 00 S 35 00 E
Emlembe Mountain . . . . 25 55 S 31 07 E
Epupa, see Monte Negro Falls
Erongo Mountains . 21 40 S 15 40 E
Etosha National Park . . . . . 19 00 S 16 00 E
Etosha Pan, *salt flat* . . . . . 18 50 S 16 20 E
False Bay, see Vals Bay
Fish (Vis), *river* . . . . 28 08 S 17 11 E
Fish (Vis) River Canyon . . . . 27 40 S 17 35 E
Fria, Cape . . . . . 18 27 S 12 01 E
Gams, Mount . . . . 23 21 S 16 13 E
Gariep (Hendrik Verwoerd) Dam . . 30 38 S 25 30 E
Golden Gate Highlands National Park . . . 28 30 S 28 37 E
Good Hope, Cape of . . . . . 34 21 S 18 28 E
Gorongosa National Park . . . . 18 45 S 34 20 E
Great Dyke Hills . . . 18 00 S 30 30 E
Great Escarpment . . . . 32 00 S 18 30 E
Great Fish (Groot-Vis), *river* . . 33 29 S 27 08 E
Great Karas Mountains . . . . 27 20 S 18 45 E
Great Karoo (Central Karoo, or Groot Karoo) Upland . . . 33 00 S 22 00 E
Great (Groot) Kei, *river* . . . . . 32 41 S 28 22 E
Great Swart Mountains (Groot Swartberg) . . . . . 33 25 S 22 30 E
Great Zimbabwe National Monument . . . . . 20 17 S 30 56 E
Groot-Vis, see Great Fish
Hendrik Verwoerd Dam, see Gariep Dam
Hexrivier, Mount . . . 29 48 S 26 19 E
Highveld, *plateau* . . 19 00 S 30 55 E
Highveld, *plateau* . . 29 00 S 25 00 E
Huib-Hoch Plateau . . . . . 27 10 S 16 50 E
Huila Plateau . . . . . 15 20 S 14 50 E
Hwange (Wankie) National Park . . . . 19 10 S 26 30 E
Incomati, see Komati
Indian Ocean . . . . . 30 00 S 37 00 E
Inyangani, Mount . . 18 18 S 32 51 E
Iona National Park . . . . . 16 40 S 12 20 E
Joubert Mountains . . . . . 18 25 S 13 55 E
Kaap Plateau . . . . . 27 30 S 23 45 E
Kafue, *river* . . . . . 15 56 S 28 55 E
Kafue Flats, *plain* . . 15 40 S 27 25 E
Kafue National Park . . . . . 14 30 S 26 10 E
Kalahari Desert . . . . 23 00 S 22 00 E
Kalambo Falls . . . . 08 36 S 31 14 E
Kambondo, *river* . . 13 03 S 23 29 E
Kamies, Mount . . . . 30 20 S 18 07 E
Kaokoveld, *plateau* . . . . 20 00 S 14 00 E
Kariba Dam . . . . . 16 31 S 28 46 E
Kariba, Lake, *reservoir* . . . . . 17 00 S 28 00 E
Kasungu National Park . . . . . 12 55 S 33 07 E
Katumbela, see Catumbela
Kaulashishi Hill . . . . 15 20 S 29 46 E
Kayamba Hills . . . . 14 10 S 26 57 E
Kgalagadi Transfrontier Park . 25 41 S 20 20 E
Khami Ruins National Monument . . . . . 20 08 S 28 23 E
Khomas Highland . . 22 40 S 16 20 E
Kissama, see Quiçama
Komati (Incomati), *river* . . . . . 25 46 S 32 43 E
Krokodil (Crocodile), *river* . . 24 12 S 26 53 E
Kruger National Park . . . . . 23 50 S 31 30 E
Kuiseb, *river* . . . . 23 07 S 14 30 E
Kunene (Cunene), *river* . . . . . 17 16 S 11 47 E
Kwando (Cuando), *river* . . . 16 32 S 22 07 E
Kwango (Cuango), *river* . . . 16 57 S 16 06 E
Kwanza (Cuanza, or Quanza), *river* . . . . 09 21 S 13 09 E
Lake Malawi National Park . . . . 14 05 S 34 55 E
Lebombo Hills . . . . 26 15 S 32 00 E
Lichinga Plateau . . . 13 16 S 35 15 E
Ligonha, *river* . . . . . 16 51 S 39 09 E
Likoma Island . . . . . 12 04 S 34 44 E
Limpopo, *river* . . . . 25 12 S 33 32 E
Limpopo Valley . . . 23 00 S 29 00 E
Linyanti (Chobe), *river* . . . . . 17 47 S 25 10 E
Lipobane Point . . . . 16 56 S 39 07 E
Little Karoo Upland . . . . . 34 10 S 21 10 E
Lochinvar National Park . . . . . 15 55 S 27 15 E
Lourenço Marques, see Delagoa Bay
Lowveld, *plateau* . . 21 30 S 31 15 E
Luando Integral Nature Reserve . . 11 06 S 17 40 E
Luangwa (Aruângua), *river* . . 15 36 S 30 25 E
Luapula, *river* . . . . . 09 26 S 28 33 E
Lugenda, *river* . . . . 11 26 S 38 33 E
Lukanga Swamp . . 14 25 S 27 45 E
Lundi, see Runde
Lungué-Bungo (Lungwebungu), *river* . . . . . 14 19 S 23 14 E
Lúrio, *river* . . . . . 13 31 S 40 32 E
Lusutfu, see Maputo
Mafinga Hills . . . . . 10 00 S 33 20 E
Mahoni, Mount . . . . 10 28 S 33 40 E
Makgadikgadi (Makarikari) Pans, *depression* . . . . 20 40 S 25 40 E
Makgadikgadi Pans Game Reserve . . . . . 20 30 S 24 40 E
Malanje Highlands . . . . . 08 45 S 16 00 E
Malawi, see Nyasa, Lake
Malombe, Lake . . . 14 38 S 35 12 E
Mana Pools National Park . . . . 16 00 S 29 25 E
Maputo (Lusutfu, or Usutu), *river* . . . 26 11 S 32 42 E
Maravia Highlands . . . . . 15 00 S 31 30 E
Marico (Groot-Marico), *river* . . . . . 24 12 S 26 53 E
Matabele Plain . . . . 16 20 S 23 10 E
Matopo (Matopos) Hills . . . . . 20 35 S 28 40 E
Messalo, *river* . . . . 11 40 S 40 26 E
Moçâmedes, see Namib Desert
Môco, Mount . . . . 12 28 S 15 11 E
Mogalakwena, *river* . . . . . 22 27 S 28 55 E
Mohokare, see Caledon
Molopo, *river* . . . . . 28 31 S 20 13 E
Monte Negro (Epupa) Falls . . . . 17 00 S 13 15 E
Morupule, *coal mine* . . . . . 22 32 S 27 07 E
Mountain Zebra National Park . . . . 32 15 S 25 25 E
Mozambique Channel . . . . . 20 00 S 43 00 E
Mozambique Plain . . . . . 20 30 S 35 00 E
Mozambique Plateau . . . . . 15 00 S 37 00 E
Msandane Hill . . . . 21 37 S 28 58 E
Muchinga Escarpment . . . 14 35 S 29 30 E
Muchinga Mountains . . . . . 12 00 S 31 45 E
Mufulwe Hills . . . . 13 55 S 29 45 E
Mwenezi (Nuanetsi), *river* . . 21 25 S 30 44 E
Mweru, Lake . . . . 09 00 S 28 45 E
Mweru Wantipa, Lake . . . . . 08 42 S 29 46 E
Namib (Moçâmedes) Desert . . . . . 23 00 S 15 00 E
Namib-Naukluft Park . . . . . 23 05 S 15 10 E
Namúli, Mount . . . . 15 21 S 37 00 E
National West Coast Recreation Area . . . . . 21 50 S 14 10 E
Neve, Serra da . . . . 13 43 S 13 10 E
Ngami, Lake . . . . . 20 30 S 22 40 E
Niassa National Park . . . . . 12 00 S 37 00 E
North Luangwa National Park . . . . 11 55 S 32 05 E
Nossob, *river* . . . . . 26 54 S 20 41 E
Notwani, *river* . . . . 23 44 S 26 57 E
Ntlenyana, Mount . . 29 28 S 29 16 E
Ntwetwe Pan . . . . . 20 35 S 25 30 E
Nuanetsi, see Mwenezi
Nuweveld Range . . 32 10 S 22 20 E
Nyasa (Malawi, or Niassa), Lake . . . 12 00 S 34 30 E
Nyika National Park . . . . . 10 48 S 33 48 E
Nyika Plateau . . . . . 10 39 S 33 30 E
Okavango (Cubango), *river* . . 16 50 S 22 24 E
Okavango (Okovanggo) Swamp . . . . . 19 30 S 23 00 E
Okwa, *river* . . . . . 22 26 S 22 58 E
Olifants (Elefantes), *river* . . 24 03 S 32 40 E
Omatako, *river* . . . . 21 12 S 16 13 E
Orange, *river* . . . . . 28 38 S 16 27 E
Otavi Mountains . . . 19 35 S 17 36 E
Phofung, see Sources, Mont aux
Púngoè (Pungwe), *river* . . . 19 50 S 34 48 E
Quanza, see Kwanza
Quiçama (Kissama) National Park . . . . 09 45 S 13 35 E
Revuè, *river* . . . . . 19 50 S 34 02 E
Roggeveld Mountains . . . . . 31 50 S 19 50 E
Rovuma (Ruvuma), *river* . . . 10 29 S 40 28 E
Ruacana Falls . . . . 17 24 S 14 13 E
Runde (Lundi), *river* . . . . . 21 19 S 32 24 E
St. Helena Bay . . . . 32 45 S 18 05 E
Sand, *river* . . . . . 22 18 S 30 07 E
Santa Maria, Cape . . . . . 13 25 S 12 32 E
São Sebastião Point . . . . . 22 05 S 35 24 E
Sapi Safari Park . . . 15 50 S 29 43 E
Sapitwa Peak . . . . . 15 57 S 35 36 E
Save, *river* . . . . . 21 00 S 35 02 E
Schwarz Mountains . . . . . 26 00 S 17 05 E
Sehlabathebe National Park . . . . 29 52 S 29 06 E
Shire (Chire), *river* . . . . . 17 42 S 36 19 E
Shirwa, see Chilwa
Skeleton Coast Park . . . . . 20 00 S 13 20 E
Sneeu Mountains . . 32 01 S 24 13 E
Sofala Bay . . . . . 20 11 S 34 45 E
Sources, Mont aux (Phofung) . . . 28 46 S 28 52 E
South Luangwa National Park . . . . 13 00 S 31 40 E
South Rukuru, *river* . . . . . 10 44 S 34 14 E
Springbok Flats . . . 24 45 S 28 50 E
Stormberg Range . . 31 25 S 26 50 E
Sucoma Estates, *point of interest* . . 16 13 S 34 52 E
Sundays, *river* . . . . 33 43 S 25 51 E
Swakop, *river* . . . . . 22 41 S 14 31 E
Tanganyika (Tanganika), Lake . . . . . 06 00 S 29 30 E
Tchevira, Mount . . . 14 19 S 13 53 E
Tigres, Bay of . . . . 16 40 S 11 47 E

Timbué Point ..... 18 48 S 36 22 E
Tsodilo Hills ..... 18 45 S 21 46 E
Tugela, *river* ..... 29 14 S 31 30 E
Tuli, *river* ..... 21 48 S 29 04 E
Tuli Block Farms .. 22 50 S 28 00 E
Ugab, *river* ..... 21 11 S 13 37 E
Umfoiozi Game Reserve ..... 28 20 S 31 50 E
Umvukwe Range .. 17 20 S 30 38 E
Usutu, see Maputo
Vaal, *river* ..... 29 04 S 23 38 E
Vals (False) Bay ... 34 15 S 18 35 E
Van der Kloof Dam ..... 30 00 S 24 45 E
Vavele, Serra ..... 11 55 S 15 30 E
Verneukpan, *salt flat* ..... 30 00 S 21 05 E
Victoria Falls ..... 17 55 S 25 21 E
Victoria Falls National Monument ..... 17 55 S 25 21 E
Viphya Mountains .. 11 50 S 33 46 E
Vis, see Fish
Vis River, see Fish River Canyon
Wankie, see Hwange National Park
Water Mountains .. 24 20 S 28 15 E
Wilge, *river* ..... 27 07 S 28 24 E
Willem Pretorius Game Reserve ... 28 19 S 27 16 E
Xau (Dow), Lake, *pan* ..... 21 15 S 24 40 E
Zambezi (Zambeze), *river* .. 18 50 S 36 17 E
Zambezi Escarpment ..... 15 30 S 29 00 E
Zomba Massif .... 15 20 S 35 18 E

exposed granitic intrusions in the complex give rise to a series of resistant hills.

Erosion surfaces

The level to gently undulating plains in interior southern Africa are known as erosion surfaces. The breakup of Gondwana gave rise to a series of new erosion base levels as periodic regional uplift took place. Streams eroded down to these base levels, giving rise to plains that are still seen throughout the region today and that are known as the African erosion surface. At the end of the Miocene (23.7 to 5.3 million years ago) renewed tectonic uplift and tilting again created new base levels. Rejuvenated fluvial activity at this time cut down into previously weathered surfaces exposing resistant hills (*bornhardts*) and piles of partially weathered rocks (castle kopjes). The new plains formed at this time are known as the Post-African I erosion surface.

Dissecting the northeastern highlands and plateaus are the rift valleys of the Zambia-Niassa Trough. These were formed by a combination of pre-Karoo folding and post-Karoo faulting and are an extension of the East African Rift Valley. The main trough floor lies 1,000 to 3,300 feet below the adjacent plateaus. The best-developed rift valley in this region is occupied by Lake Nyasa (Malaŵi). It is bounded by the Kipengere Range and Livingstone Mountains in Tanzania and by the Nyika Plateau and Viphya Mountains in Malaŵi. Several troughs branch from it, such as along the Shire valley and the Gwembe-Lungwa trough.

Gerald Cubitt

Whaleback outcroppings of Precambrian granite of the basement complex in the Matopo Hills, southwestern Zimbabwe.

Flanking the interior plateaus from southern Mozambique to northern Angola lies the Marginal Zone, which mainly comprises well-dissected rolling topography, although the extreme south is dominated by the basin-and-range topography of the Cape Ranges. These mountains were formed during the Permian (286 to 245 million years ago); the valleys deepened during the Triassic and Jurassic periods, forming the present landscape by the mid-Cretaceous. An important series of arid basins known as the Great Karoo lie at altitudes of 1,000 to 2,600 feet between the Great Escarpment and the Cape Ranges. In Namibia the coastal margin includes the Namib (Moçâmedes) desert and consists of four geomorphological regions: ridged topography (trough Namib), debris-covered surfaces (plain Namib), dune fields (dune Namib), and an interior range of well-dissected table mountains (Kaokoveld).

The coastal plains of Mozambique expand from a narrow strip in the north to a wide plain south of the Save River. The combination of low relief and runoff from the high ground to the west causes a considerable amount of flooding in the lagoons and swamps found on these plains. The coastal plain in northern Angola exhibits similar topography, although the south is occupied by the Namib desert.

The Zambezi and Limpopo rivers

Drainage. The region is generally drained eastward toward the Indian Ocean, a pattern exemplified by the largest rivers, the Zambezi and Limpopo. Some 2,200 miles in length, the Zambezi is the longest river in the region, and its catchment includes significant parts of Angola, Zambia, and Zimbabwe. Important dams at Kariba and Cabora Bassa restrict much of its flow. The 1,000-mile-long Limpopo, immortalized by the British author Rudyard Kipling as "the great grey-green, greasy Limpopo River, all set about with fever trees," drains large areas in Botswana and South Africa. Other important rivers draining eastward are the Lúrio, Pondo, and Save.

The only major river flowing into the Atlantic Ocean is the Orange (1,300 miles long), which drains parts of South Africa, Lesotho, and Namibia. The Orange is extensively used to provide irrigation water in South Africa, and the scale of withdrawals is so high that in some years the river barely reaches the sea. Two other rivers, the Kunene (Cunene) and Kwanza (Cuanza) in Angola, also drain westward.

The most interesting large-scale pattern is the radial drainage of the Bié (Angolan) Plateau, with tributaries of the Congo (Zaire) River heading northward, those of the Zambezi westward and southwestward, those of the Okavango (Cubango) and Kunene southward, and those of the Kwanza westward. Noteworthy is the Okavango River, which rises on the Bié Plateau, heads into the northern reaches of the Kalahari desert, and drains into the swamplike Okavango delta, never reaching the sea. Two further important internal drainage systems are found in the Kalahari Basin—the Etosha Pan in the northwest and the Molopo-Nossob system in the south.

Soils. A wide range of soils are found in southern Africa. They vary from the largely infertile and poorly developed soils found in the deserts and on higher mountain slopes to old soils with deep, well-developed profiles on some of the northern plateaus. This great diversity results from the many significant local variations in the region's climate, lithology, and topography.

In the arid parts of the region, the soils either show little development or are characterized by soluble salt accumulations in their topsoils. Soil types are differentiated largely on the basis of their mineral composition. Calcisols are rich in calcium carbonate, gypsisols in gypsum, and solonchaks and orthic solonetz are often rich in halite. On the drier semiarid plateaus a larger fraction of the soil texture is clay. Somewhat more fertile planosols, vertisols, and cambisols are common. In wetter areas and on the northern plateaus the most common soils, ferralsols and luvisols, are found. Dominated by the movement of humus, minerals, and clays from the topsoil to subsoil, these soils are the most widespread in southern Africa. Sand-

rich arenosols have their maximum development on the Kalahari Sands in Angola and Zambia, and nutrient-poor, sand-rich regosols cover large parts of Angola. Histosols, rich in organic matter, are found locally in Malawi and Zambia.

The fertility of many of the soils in arid and semiarid areas is low because of the lack of organic matter in the topsoil and the slow rates of breakdown resulting from a lack of moisture. In most areas, the agricultural exploitation of such soils depends on irrigation, although this can lead to salinization and waterlogging. Luvisols are perhaps the most fertile soils in the region; they are extensively used by both large-scale commercial and small-scale peasant farmers. Ferralsols are generally infertile because their age and location on the high plateaus have resulted in the loss of most of their nutrients by leaching. Many of the region's soils are prone to water and wind erosion when cultivated; soil erosion by water is severe in some parts of Zimbabwe, Swaziland, Lesotho, and South Africa.

Climate. Southern African climates are seasonal, ranging from arid to semiarid and from temperate to tropical. The seasonality is an important control on plant growth and a regulator of river flows. Moreover, it affects soil and landform development through its controls on rock weathering and water and wind erosion.

Effects of seasonality

In the summer (October to March), a prolonged low-pressure system develops over the region, drawing in moist air from the surrounding oceans. The converging moist air rises and cools, giving rise to intense rainstorms. In the winter (April to September), low-pressure areas move northward, and the influx of high pressure over the region leads to stable, dry conditions, although wet, westerly winds continue to bring rain to the southern Cape region. The lack of cloud cover over most of the region in winter leads to ground frosts, and these, combined with very low humidities in some areas, cause black frosts, which damage the more vulnerable broad-leaved plants.

The climatic seasonality creates four climatic zones. Western Cape province has a warm temperate Mediterranean climate with cool, wet winters and hot, dry summers. The mean annual rainfall at Cape Town is 20 inches (510 millimetres), most of which falls between April and September; mean daily maximum temperatures reach 79° F (26° C) in February, the warmest month, and drop to a low of 63° F (17° C) in July. The interior plateaus have a semiarid steppe climate characterized by great extremes of temperature between the winter and summer months. In the Highveld to the south the climate is temperate, but at lower altitudes and farther north the climate varies between tropical and subtropical. Most rain (28 inches at Johannesburg and 33 inches at Lusaka) falls in summer. The winters are dry, with a high diurnal temperature range. Mean daily maximum summer temperatures range from 75° to 78° F (24° to 26° C) at Johannesburg and from 78° to 88° F (26° to 31° C) at Lusaka. The mean daily maximum temperatures in winter range from 62° to 77° F (17° to 25° C) at both stations, with Lusaka being slightly warmer on average. The eastern part of the region has a subtropical to temperate maritime climate with high year-round precipitation due to the orographic effect of the Great Escarpment on easterly winds. The rainfall amounts are greatest in the summer, however. At Maputo on the Indian Ocean coast, the mean annual rainfall is 30 inches, and the mean temperatures range from 75° F (24° C) in July to 88° F (31° C) in February. The Namib desert has a tropical arid climate with very little rainfall. The mean amount of precipitation at Namibe (formerly Moçâmedes) on the Angolan coast, for example, is about two inches, and at Walvis Bay, farther south along the Namibian coast, it falls to only one inch. Frequent condensation from coastal fog (known as *caicimbo* in Angola) does occur, however, providing some additional limited moisture. These fogs roll inland off the South Atlantic and are formed when moist air cools to its dew point and condenses as it passes over the cold surface waters of the Benguela Current.

Droughts are common in much of southern Africa, and there is evidence that an 18-year cycle of drought occurs in parts of the region. Drought was widespread in the 1960s, and the early 1980s saw droughts in Botswana, South Africa, and Zimbabwe. Their effects were mainly restricted to semiarid areas.

Plant life. Four main types of vegetation are found in southern Africa. The north is dominated by savanna woodlands known as miombo forest. To the south of these are a series of dry woodlands. Arid and semiarid grassland, shrubland, and bushland dominate the Namib and Kalahari deserts and their environs. Mediterranean vegetation is found along the southern coast.

Woodlands

Miombo woodlands stretch from the Bié Plateau through Zambia and Malawi into northern and central Mozambique and south into Zimbabwe. They are characterized by a 20- to 100-foot-high single story of tree canopy that is open enough to allow a grassy understory. Deciduous species of the genera *Brachystegia* and *Julbernardia* dominate the overstory.

To the south of the miombo forest is a belt of dry woodlands stretching from southern Angola and northern Namibia into southern Mozambique and Swaziland. Most extensive are the *Acacia*-dominated woodlands. In southeastern Angola, western Zimbabwe, and Zambia, *Baikiaea* woodlands, dominated by the Zambezi teak (*Baikiaea plurijuga*), and wild seringa (*Burkea africana*) savannas have developed on the Kalahari Sands. Mopane woodland, dominated by *Colophospermum mopane*, is another distinct type of dry woodland. It is common in southern Angola and Botswana, although it extends as far north as southern Malawi. In the Kalahari, woodlands dominated by *Acacia* species and *Boscia albitruncata* are well developed.

Arid and semiarid vegetation dominate the Namib and Kalahari and the areas surrounding them. The two most extensive vegetation types are the veld grasslands (and, less commonly, wooded grasslands) and the succulent, or sclerophyllous, shrublands. The former are most common in the high continental interior (*e.g.*, the Transvaal region), whereas the latter dominate South Africa's Cape provinces and southern Namibia.

Mediterranean vegetation is found along the mountainous southern and southeastern coast. Diverse and luxuriant temperate woodlands are found in a number of moister sites, but woodland and thickets (macchia, or *fynbos*) of evergreen sclerophyllous species are more common. Species of *Protea* are found in these areas.

Of limited areal extent and only locally important are the subalpine and alpine grasslands and heathlands found in Lesotho and surrounding parts of South Africa. High altitudes and cooler temperatures also give rise to the unique coniferous forests of the Drakensberg.

Deforestation is mainly caused by land clearance for agriculture, the collection of domestic fuelwood, and the felling of timber for construction. The main areas of deforestation are along the "Line of Rail" in Zambia, central and southern Malawi, on the Highveld and in the Limpopo valley in Zimbabwe, in southern Mozambique, in southwestern Angola and adjoining northwestern Namibia, in Swaziland, and in parts of northern and eastern South Africa. Reforestation is generally on a small scale throughout the region, although there are important conifer plantations in South Africa and Swaziland.

Animal life. The semiarid plains and plateaus that cover much of the region contain animals commonly associated with the East African plains—*e.g.*, antelope, gazelle, elephant, and the big cats. However, different animals are found in the coastal woodlands of South Africa and in the desert regions to the north and northwest. Many habitats have been extensively modified by agriculture, thus restricting the ranges of certain species that were formerly more widespread.

Endangered species

Most of the endangered species (*i.e.*, those under immediate threat of extinction) in southern Africa are mammals, especially large mammals, but, if vulnerable species are included, some reptiles are also threatened. Endangered species include the giant golden mole (*Chrysopalax trevelyani*), riverine rabbit (*Bunolagus monticularis*), and Cape mountain zebra, all of which are found only in South Africa; the black-faced impala in Namibia; the sable antelope from Angola; and the black rhinoceros, which is

mainly restricted to Zimbabwe. The only endangered bird is the Cholo alethe found in Malawi.

Vulnerable mammals include another species of zebra, the Hartmann's mountain zebra of Namibia, as well as the cheetah, brown hyena, lechwe, African elephant, African wild dog, and bontebok. The Smith's dwarf chameleon and the geometric tortoise, found in the Cape region, are vulnerable species of reptile, as is the Nile crocodile, found along the Zambezi. The Cape platanna, an aquatic frog, is also vulnerable.

Cape mountain zebra (*Equus zebra zebra*) number fewer than 500 and are almost entirely restricted to the Mountain Zebra National Park in South Africa. The geometric tortoise (*Psammobates geometricus*) is found only in the Cape region. Its near extinction has been brought about by the destruction of its habitat by agricultural and urban expansion—less than 5 percent of its natural range remains. The best-known vulnerable and endangered species are the African elephant (*Loxodonta africana*) and black rhinoceros (*Diceros bicornis*). In both cases, hunting and poaching—for ivory in the case of elephants and potions and dagger handles (popular in Yemen) in the case of the black rhinoceros—have significantly reduced their numbers.

Conservation measures are variable in their application. Namibian wildlife officers have begun to saw off rhino horns to make them less attractive to poachers, although most southern African countries oppose the ban on the ivory trade. There are 16 large national parks and game reserves in the area, as well as many smaller ones. Most of them are located in the open or partially wooded plains, the most famous being South Africa's Kruger National Park. (A.C.Mi.)

THE PEOPLE

Contemporary ethnicity and class structure were forged in southern Africa within a crucible of interacting ethnic units, not through a process of timeless, apartheid-like separation of one cultural "tradition" from another. Many of these transformations were indigenously inspired and followed upon the introduction and adoption of agricultural and pastoral economies and metallurgy at the beginning of the 1st millennium AD. Others came about as responses to new opportunities and conditions originating outside the continent, first with the introduction of intercontinental trade at the end of the 1st millennium and, later, with European colonization. In the first instance, principles of kinship, residence, and tenure were flexibly adjusted to accommodate changing political and economic opportunities. In the second, a 20th-century ethnic consciousness was constructed using past history, language, and culture as the building blocks from which to create a new ideology.

Language groups. The peoples of southern Africa today can be divided into speakers of two language families, Khoisan and Bantu. The linguistic differences separating these families are great, attesting to a long history of separate development before their present juxtaposition south of the Zambezi River. In earlier times, Khoisan languages were spoken over most of eastern and southern Africa. They have now been displaced in many areas by Bantu languages and, in parts of Namibia and in the Cape region of South Africa, by European languages.

Khoisan. Taking each family separately, Khoisan peoples can be further subdivided into San and Khoi groupings. ("Khoi" is an ethnic term and "Khoe" is linguistic; that is, Khoi peoples speak Khoe languages.) In former times the pejorative terms "Bushmen" and "Hottentots" were often used to refer to these sub-groupings: "Bushmen" if they lived by hunting and gathering; "Hottentots" if they were herders. The diversity encompassed by languages of this family is a testament to its great antiquity in southern Africa. San languages include both Northern (Zhu and !xũ) and Southern (of which only /nhuki, //xegwi, and !kò are extant) variants. A wedge of Khoe languages spoken in the central Kalahari separates the Northern from the Southern San. Today only a few Khoisan speakers hunt and gather for substantial portions of their livelihood, mainly on the northern and southern flanks of the Kalahari desert in Botswana and Namibia. Even there they supplement their diet with domestic grains and milk while tending herds of cattle, goats, and sheep.

Click languages

A characteristic of Khoisan languages is their widespread use of click sounds as consonants (more than 70 percent of San words begin with a click). Each click has a distinctive set of accompaniments (effluxes) that combine to produce as many as 85 distinct click segments, making these languages among the world's most phonetically complex. Although commonly grouped into a single language family, the linguistic differences separating Khoe and San are great, and some scholars dispute the merger. They argue that this results in an artificial grouping of historically unrelated people. Such linguistic divergence is not totally unexpected, however, and may simply reflect thousands of years of social differentiation occurring among small, isolated hunting bands.

Khoe speakers include people in southern Angola and the Okavango delta (Kwadi, Kxoe, Buga, //ani), the central Kalahari (G/wi, G//ana, Naro), and the Khoikhoi speakers of Namibia (Nama) and the Cape provinces of South Africa (!ora). Some linguists would include the Tanzanian languages of Sandawe and possibly Hadza as distant relatives of the Khoe subgroup. If accepted, this would serve to extend the prehistoric distribution of this group well northward into East Africa. In contrast with San hunters, Khoi were thought of as the archetypal pastoral nomads of southern Africa even though some, such as those in the central Kalahari, were labeled as "Central Bush" because they were living as hunters and gatherers at the beginning of the 20th century. That this was not always the case is indicated by linguistic data that traces the origins of Khoi herding to this same region approximately 2,500 years ago. From this area they spread southward, reaching as far as the Cape by the 1st century AD. Ceramics and remains of domesticated sheep found in Late Stone Age contexts there indicate that an indigenous transformation from foraging to pastro-foraging (herding and foraging) had already begun before the first Bantu farmers arrived.

South of the Limpopo, Khoisan herders were the probable source of cattle and sheep for newly arrived Bantu farmers—not the reverse. The use of Khoisan-derived terms for these domesticated animals is universal among southeastern Bantu peoples such as Zulu, Xhosa, Sotho, and Tswana. This points to a long and stable history of relations between these peoples and Khoisan pastro-foragers in the region. Except in the case of 17th- and 18th-century northward migrations of southeastern Bantu speakers, these terms are not used north and east of the Kalahari.

Additional evidence for long-term interaction is provided by the click sounds incorporated in many Bantu languages, including Yei, Kgalagadi, Sotho, Zulu, Ndebele, and Xhosa. These are distributed from north to south following the 20-inch-rainfall isohyet that divides the subcontinent into agricultural and pastro-foraging economic zones. West of this line, conditions favoured the formation of convergent economies that utilized indigenous wild species adapted to these drier environments and livestock that could be seasonally moved to follow the rains and changing availability of pasturage. Compound economies formed, with autochthonous (indigenous) foragers and food producers coexisting with, and complementing, Bantu economies up to the present day. Because of low and unreliable rainfall, hunting and gathering was more heavily, or at least more obviously, relied upon than it was farther east. Considerable evidence for genetic admixture also occurs along this line, with some groups, such as the Deti and Tshukhoe in central Botswana and the Damara (Bergdama), Hain//um, and Kwadi of northern Namibia, being dually classified as genetically Bantu but linguistically Khoi. While some herders of Khoi ancestry still live in parts of Namibia, many of them now speak Afrikaans (a European language derived from Dutch) rather than Khoisan and live like Afrikaner farmers. Khoi peoples have almost entirely disappeared from the Cape area, most being absorbed into other ethnic elements, particularly the mixed-race ("Coloured") population.

Bantu. The large linguistic differences separating languages of the Khoisan family are not paralleled to the

same extent in the Bantu family, even though members of the latter are much more widely dispersed across central and southern Africa. This reflects the recentness of their divergence from a common ancestral language. Originally, proto-Bantu was spoken only in a small area north of the tropical forests in the grasslands of present-day Cameroon. About 3,000 years ago, however, an expansion of Bantu speakers began that resulted in their covering most of Africa south of the Equator by the 3rd or 4th century AD. The reasons for this expansion are presently unknown, though it could be related to the acquisition of new technologies, such as pot making, or new crops, such as sorghum, millet, and perhaps bananas, that enabled them to more efficiently exploit the forest margins. This initial migration took place at a "Neolithic" stage of development; knowledge of metalworking had not yet penetrated south of the Sahara. Reconstruction of proto-Bantu vocabulary shows that these people made pottery and had begun to cultivate, or at least intensively exploit, indigenous oil palms and yams before their expansion southward. They also kept goats but probably not cattle.

Spread of Bantu languages

Traditional cultures. *The spread of Eastern Bantu.* Early in their history, a division of Bantu into two major branches—Eastern and Western—took place. Some authors put this split at the very beginning of Bantu expansion from Cameroon. Others suggest it occurred later, and much farther south, in the area centred on the lower reaches of the Congo River. The earliest remains of mixed-farming settlements in the area immediately west of Lake Victoria date to the 2nd and 3rd centuries BC. These sites provide the first indications of the arrival of Eastern Bantu in the region and contain remains of iron smelting as well as distinctively decorated pottery. Sorghum and millet, the principal domestic crops, were acquired through contact with farmers living along the grassland savannas north of the forest after Eastern Bantu diverged from Western Bantu. Cattle were kept, perhaps adopted through contact with Central Sudanic herders in the region to the west and south of the upper Nile.

Routes of Eastern Bantu expansion

By the 1st century AD, Eastern Bantu had spread southward around the East African Rift Valley to the Kenya coast. Related settlements dating from the 3rd and 4th centuries line the Indian Ocean as far south as KwaZulu/Natal in South Africa. Ceramics and evidence of ironworking are common on these coastal sites, but no remains of domestic stock have been found. Instead, shellfish were relied upon as a source of protein.

A second route of Eastern Bantu expansion lay inland, along a course between Lake Nyasa and the eastern Kalahari. Along this route, tsetse-fly-infested areas could be avoided by Iron Age agro-pastoralists, giving them an opportunity to bring cattle across the central Zambezi onto the high plateaus of Zimbabwe and Northern Transvaal. Evidence for cattle-keeping farmers on the Highveld bordering the Limpopo River becomes abundant after the 4th century AD.

Today, Eastern Bantu-speaking societies south of the Zambezi include the Shona and Tonga of present-day Zimbabwe and Mozambique, the Northern Nguni (Swazi and Zulu) and Southern Nguni (Pondo, Tembu, and Xhosa), who occupy the coastal Lowveld of KwaZulu/Natal, and Sotho-Tswana peoples who are found on the southern African Highveld from the Transvaal region through eastern Botswana to the southern edge of the Okavango delta.

Western Bantu expansion. The Western Bantu branch expanded directly southward from Cameroon through a broad area where the Eastern complex of cereal grains, cattle, and sheep were not well adapted and in which they are virtually absent today. The fact that all Western Bantu vocabulary for sorghum and millet derives exclusively from Eastern Bantu indicates this complex spread westward below the tropical forest at a later date. Cattle, in addition, seem only to have arrived west of the Okavango River in southern Angola about the middle of the 1st millennium AD. Lexical data from Western Bantu languages such as Herero suggest that they acquired these animals directly from Eastern Bantu sources. Between the Okavango and the upper Zambezi, however, the situation is more complex, and evidence there indicates that indigenous Khoisan foragers, fishers, and cattle nomads were already present when Bantu speakers first arrived. West of the Okavango, it is uncertain whether domestic cattle and sheep passed independently to Khoisan and Western Bantu or whether the use of these animals spread as a single complex together with cereal grains. The distribution of terms for a number of varieties of banana provides indications for yet another early diffusion of cultivars from east to west, this time perhaps through the forest rather than around its southern edges. Fishing, foraging, oil palms, and shellfish and other aquatic resources contributed significantly to the early Western Bantu diet. Goats, cowpeas, and indigenous yams were also important.

Early ironworking

The earliest evidence of ironworking south of the Sahara dates to the 6th century BC in the upper Nile and to the 5th century BC in Nigeria. By the 3rd century, iron was being worked as far south as Gabon and the Congo. Below the Equator, dates from the west coast are coeval with those from sites in East Africa. As a result, the possibility cannot be ruled out that iron smelting diffused southward to Neolithic Bantu along two separate routes: one eastward from sources in the upper Nile and the other southward from early centres in Nigeria or the western Sahel. In addition to its productive use for tools such as hoes, axes, knives, and spears, iron (and copper) functioned as luxury trade goods, creating and sustaining intercommunity relations across the region. Once the techniques of metallurgy were known, they spread quickly through sub-Saharan Africa, reaching as far south as the Transvaal region and KwaZulu/Natal by the 3rd or 4th century AD.

The acquisition of cereal crops, domesticated cattle, and sheep facilitated the expansion of Western Bantu speakers into the drier savannas of southern Angola and western Zambia. Initially, the network of southward-flowing rivers provided the main avenues of inland movement; the Atlantic coast also was exploited. Exchange of salt, fish, and probably metals kept the peoples of these different environments in contact. Metalworking, in particular, was of great importance to the developing political economy of the Angolan region. Metal served as a commodity of exchange, a store of wealth, and a symbol of chiefly authority. Over time, the Western group differentiated into Umbundu and, farther south, into the Bantu languages of southern Angola, northern Namibia, and northwestern Botswana: that is, into present-day Ambo, Herero, Kwangari, Nyaneka, Ndombe, Ngonyelu, and Ngumbi. Middle Zambezi peoples, including the Mbukushu, Subia, Koba, and Tonga, also are related to this group.

The diversification of farming communities in southern Africa. After the arrival of Bantu farmers in the 3rd and 4th centuries AD, a period of radiation and adaptation to the varied environments of southern Africa ensued. New ecological relationships were established as fields were cleared and planted and as domesticated animals increased in number to compete with indigenous fauna for grazing and water. Iron and copper were prospected for, mined, and smelted, and larger populations settled in more permanent villages than was common for earlier foragers. A combination of local movements by farmers seeking new fields and pastures and the adoption of new technologies by indigenous hunters and gatherers resulted in a complex intermingling of peoples and economies below the Zambezi, Okavango, and Kunene rivers. A dichotomization of southern African peoples into foragers and farmers—Khoisan and Bantu—oversimplifies what was a more synergistic regional political economy. At the time of Bantu entry into the region south of the Zambezi 2,000 years ago, there was thus less material differentiation among peoples than was to appear later; progressive linguistic and ethnic diversification was a feature of the political transformations of the 2nd millennium AD. Overall, there is little evidence for a rapid domination of the subcontinent by these newcomers.

Effects of rainfall on culture

Rainfall had a profound influence upon the development of southern African societies, and ways of life varied directly in proportion with decreases in rainfall from east to west. The 20-inch isohyet marks the effective limit of arable land. This rainfall line cuts diagonally across the region from the Okavango delta to the central Limpopo

valley and then turns southward to run along the western side of the Drakensberg. As one moves west from this line, rainfall decreases and, in concert with this change, the contribution of livestock to the local economy increases. Soon after their introduction, cattle, goat, and sheep populations exploded in the drier portions of the subcontinent where nutritious grazing land was abundant and animal diseases were few. Such changes reflect the outcome of common ecological processes whereby rapid increases in population often occur as a matter of course when new species are introduced to colonize favourable environments. (Increases in cattle, sheep, and other animal populations following the European settlement and colonization of the American Southwest and Australia provide familiar examples of this.) In Iron Age Africa, such developments had important social and political repercussions; between AD 500 and 900 early mixed farmers exploited the potential for herd expansion offered by these rich grasslands. Archaeological sites in Botswana, Zimbabwe, and South Africa contain evidence in the form of dung deposits, some more than six feet deep, that attest to this increase in the importance of the pastoral sector of early farming communities. This was particularly true for the region west of the 20-inch rainfall line.

The impact of these changes is reflected in the layout and organization of early settlements. Typically, thatch-roofed houses constructed of closely spaced poles and sticks plastered over with a mixture of animal dung and clay were arranged in a semicircle around a central animal byre. Most such houses were between four and six feet in diameter; family compounds were composed of several such units functioning as separate sleeping, storage, and cooking areas. The deep dung deposits found at many of these Iron Age villages indicate that cattle and other domestic stock were closely watched and were probably returned to the village each evening in order to ensure their security against predators such as lions, as well as to protect them from cattle raids by neighbouring groups. Burials of men in these central cattle kraals indicate that some elements of traditional values, rituals, and beliefs can be traced back at least 1,000 years in this region. In contrast to the central location of livestock pens, grain supplies were individually held in granaries associated with each homestead or compound. Indications of communally held grain buried in dung-lined pits under the central cattle kraals have also been found. In historic times, such grain was controlled by central authorities and was used as a reserve or shelter against times of drought or warfare, when it could be dug up for redistribution.

The early development of large chiefdoms in the Limpopo valley was clearly associated with increases in cattle and their use as wealth. Both there and in neighbouring eastern Botswana a hierarchical social structure, with an associated settlement pattern based upon central towns surrounded by smaller satellite villages, had emerged by the end of the 1st millennium AD. Farther east, where rainfall is higher and more reliable, cattle herds were smaller and agriculture contributed a larger proportion of the subsistence diet. Such regional contrasts were important to the developing political economy of southern Africa. In the east, Iron Age Bantu mixed farmers gradually gained a dominant position over autochthonous foragers and pastro-foragers, ultimately subjugating, absorbing, or eliminating them. A hierarchy ensued, with foragers relegated to marginal social and economic positions. In the west different processes of interaction led not to population absorption or replacement but to an articulation of herders and foragers into multiethnic compound economies that continued into the 20th century.

The east-coast trade and political transformations. Linkages with Arab traders on the Mozambique coast were well established by the end of the 1st millennium AD. Long-distance trade was not a new innovation but simply grafted onto already existing regional networks along which cattle, salt, fish, metals, chert, ostrich eggshell beads, and other items had been flowing for many centuries. People of all ethnic backgrounds participated in the trade, funneling locally made and acquired goods to the east coast through indigenous entrepôts such as Bambandyanalo in the Limpopo valley and Toutswe and Bosutswe in eastern Botswana. By AD 900 inter-regional exchange networks stretched completely across the subcontinent from the east coast to the headwaters of the Okavango delta in southern Angola. Hunted products, including ivory, carnivore skins, and rhinoceros horn, were among the most highly valued items offered for trade. Most of these goods were probably obtained by indigenous foragers who intensified hunting activities in order to supply increased external demand for these exotic commodities; cattle, pottery, metals, and other locally produced items supplied intraregional requirements. Glass beads, marine shells, and cotton cloth were among the main items provided in exchange. These high-prestige goods served to enhance the political value of the trade to local elites who directed and controlled it.

Exchange networks

Iron and copper were extensively mined, smelted, and traded during the 1st millennium, but there is no evidence that gold was used. The 13th-century burial of an important official uncovered at Mapungubwe in the Limpopo valley—accompanied by a gold-covered statue of a rhinoceros, a golden staff, and other artifacts—is perhaps one of the earliest indications of gold mining in southern Africa. The Mapungubwe gold was panned from alluvial deposits, but, by the last half of the 2nd millennium, more than 1,000 underground mines had also been dug into the granitic bedrock of northeastern Botswana and neighbouring Zimbabwe, in some cases following ore seams down to depths of more than 100 feet. From the 13th century onward, gold was strongly associated with elite status. It was undoubtedly the most highly prized item sought by east-coast traders, taking the place formerly held by ivory. The rise of the Great Zimbabwe kingdom about 185 miles northeast of Mapungubwe is associated with this change, which, over the next 200 years, resulted in the consolidation of Shona political hegemony across the high central plateau.

With the discovery of gold, a devaluation in cattle as the principal repository of political and economic capital took place across the gold-bearing regions of northeastern Botswana and Zimbabwe. One indication of this change is suggested by the fact that livestock were no longer kept in large numbers at elite centres, as was common in former times. Instead, they were dispersed among commoners living in smaller villages and hamlets in the hinterlands. The symbolic role of the central cattle byre at earlier sites was replaced by elaborate stone walling constructed around the chiefly residences built at Great Zimbabwe and at smaller regional centres and towns from the Makgadikgadi salt pans in the west to the Mozambique coast in the east. These stone buildings served as the political centres of Shona hegemony.

In concert with the political and economic transformations that accompanied the rise to power of Great Zimbabwe, the value of trade networks reaching out to the cattle-rich areas of eastern Botswana, and beyond them to the hunting grounds of the Okavango, fell. In eastern Botswana disruption of these long-distance routes contributed to the collapse of the political hierarchies they had formerly supported. West of the Okavango, similar political and economic retrenchments occurred as the somewhat weaker economic base of emerging elites in this area was eroded and their settlements abandoned. When first encountered by Europeans toward the middle of the 19th century, Khoe-speaking and Herero (Western Bantu) herders dominated the drier central Namibian grasslands. Farther east, Mbukushu (Western Bantu) and Tswana (Eastern Bantu) mixed pastoralists were in control around the Okavango delta, with Khoi, San, Koba, Yei, and others subservient to them.

Disruption of trade networks

South of the Limpopo on the open grasslands of the Transvaal and Orange Free State, the central economic, social, and symbolic importance of livestock continued into the 2nd millennium as cattle-based chiefdoms emerged in the 15th and 16th centuries as a prelude to later Sotho and Tswana chiefdoms. Stone walling was also used extensively on these sites, but, rather than represent a break with the past, these walls provide evidence of continuity with the earlier pattern of central cattle byres surrounded by granaries and family housing compounds. Discontinu-

ities in pottery styles, however, suggest new population elements may also have moved southward into the region, becoming incorporated in the political restructuring that was spreading across southern Africa during the last half of the 2nd millennium.

Traditional agriculture

Subsistence economy. Sorghum and millet were the traditional subsistence crops of southern Africa, the soil being tilled with hand-held hoes fashioned of iron or wood. These traditional staples, while still grown in the drier regions, have now been replaced in most areas by higher-yielding corn (maize), beans, peanuts (groundnuts), and sweet potatoes of American origin. In mixed agricultural areas where cattle are present, southern Africans have also shifted from labour-intensive hand-cultivation techniques to ox-drawn plows in order to overcome the problems of low per-acre productivity and traditional peak-season labour shortages. In the past almost all households were engaged in some combination of agriculture and herding, with craft specialties such as pot making and iron smelting usually taking place during the dry season when labour was not demanded in the fields and pastures. Rights to land were vested in tribal membership, and, with adequate land resources, unequal agricultural production was based upon unequal access to labour.

Because nonfamily labour was in short supply, reciprocal work parties, polygamy, and slavery were the most common methods used to overcome peak-season labour constraints. Low population densities, simple transport systems, and poor storage facilities meant that few households had an incentive to produce a surplus much above what was needed for domestic consumption. Prior to the end of the 19th century, when the ox-drawn plow was widely adopted, discrepancies in wealth were therefore not generally based upon differences in surplus agricultural production. With the plow, wealthier families who had access to oxen could increase production. Demand for agricultural labour diminished proportionately, and it is perhaps no accident that in areas where the plow was widely adopted, such as the eastern Kalahari, polygamy disappeared soon afterward.

Cattle as a store of wealth

Cattle production was not as directly tied to labour. Their meat, milk, and skins contributed importantly to the subsistence economy; an additional function of these animals was to serve as a traditional store of wealth and a medium of exchange. Cattle were the most important source of capital investment; a man's wealth was generally estimated by the size of his herd. They were also a major source of political and economic power. Throughout the region where they were kept, institutionalized mechanisms existed through which wealthy men could gain political support by judiciously loaning out cattle to those who had few or no animals. This, and the rapid rates of herd increase experienced in the area, meant that, while returns to traditional agricultural production were low, cattle were a means to the accumulation of wealth. The increased value of oxen as draft animals, brought about by the introduction of the plow, reinforced these discrepancies. Families who lost their herds through misfortune were limited from full participation in the spheres of life that required capital and remained tied to the system through subservience and dependency to the owners of large herds.

Contemporary southern African societies. Differences in the amount and type of movable wealth from west to east across southern Africa underlay differences in the structure of social life. Variations in lineal inheritance patterns, descent, and residence after marriage, for instance, correlate with the amount of moveable wealth, particularly cattle, found in an area. In areas where they were present, cattle constituted the bride-price (*lobola; bogadi*) exchanged for a wife, her labour, and the right to all the children born. Residence after marriage was related to the amount of bride-price paid or promised. In southern Angola, northern Namibia, and northwestern Botswana, where cattle and other livestock are numerous, payment of bride-price is the norm, and residence is patrilocal (with the husband's family) or avunculocal (with the husband's maternal uncle). Among the cattle-rich societies such as the Tswana of eastern Botswana and South Africa, bride-price is uniformly high, and inheritance and descent are reckoned along patrilineal lines. Residence is normally patrilocal. Those who cannot afford to pay with cattle or other valuable goods generally perform bride service—and reside matrilocally (with the wife's family) for an agreed-upon period of time. In parts of Zimbabwe where cattle are fewer, such as among the Tawara and Zezuru, most marriages were of the second type, with prolonged matrilocal residence. In central Zambia, where inheritable wealth is almost nonexistent, matrilineal descent with bride service and matrilocal or avunculocal residence were most common.

Cattle were the most important objects of sacrifice, linking men to the spiritual world, and, in former times, they figured prominently in rainmaking and rites of passage as well. Among the Herero and neighbouring southwestern Bantu peoples, every adult male was given a set of "sacred" cows by his father and his mother's brother. These were dedicated to the spirits of deceased maternal and paternal relatives. At his death, some were sacrificed and the remainder redistributed back to designated relatives. Even the layout of southern African homesteads, with thatched houses grouped around central animal byres, reflected the symbolic and economic centrality of herding to these cultures. The bodies of important men and chiefs, ancestors who would be called upon in times of need, were buried under these central cattle kraals, reinforcing their symbolic centrality.

Kinship systems

With one major exception, the system of Iroquois kinship terminology is used throughout the region. This terminology differentiates parallel cousins (father's brothers' and mother's sisters' children) from cross-cousins (father's sisters' and mother's brothers' children), with parallel cousins usually being referred to by the same terms as those used for brother and sister. Marriage with parallel cousins is generally not permitted, but cross-cousin marriage—particularly with the mother's brothers' daughter—was often preferred. In both matrilineal and patrilineal systems, such marriages serve to reinforce linkages between lineages. In matrilineal regions, marriage to the mother's brother's daughter kept men of the same lineage together, rather than dispersing them to a number of different villages. The reverse pattern, marriage to the father's sister's daughter, is rarer, but it does occur among cattle-rich groups such as the Tswana and the Herero, where it was preferred by elites and rich commoners as a strategy to keep bride-price payments recirculating within the group.

The only exception to the use of Iroquois kinship terminology are the Shona and Thonga of Zimbabwe and Mozambique, who use Omaha kin terms. This terminology serves to set apart males linked to each other by descent through men only and is widespread among strongly patrilineal societies. In contrast with Iroquois patterns, cousin marriage is strictly forbidden, and individuals are encouraged to marry outside the local group. This is a good strategy in situations where it is desirable to create as many diverse marriage ties with strangers as possible. This was certainly the case during the period of Zimbabwean hegemony in the 15th and 16th centuries when male members of royal rank were assigned as headmen to outlying settlements according to their status. Marriage or concubinage alliances were important instruments of policy, and it is possible that the interesting variation in kinship terminology found among the Thonga and Shona between the Kalahari and the east-coast region is a testimony to these early political processes. (J.R.De.)

THE ECONOMY

The economies of those countries constituting southern Africa are highly diverse in terms of human and natural resources; the structures of these economies are also extremely heterogeneous. They have, however, shared many of the problems that have plagued the continent as a whole. These include wars, political instability, drought, fluctuations in export commodity prices, and declines in the capacity to import. As a consequence, in the 1980s many of these countries were compelled to adopt severely deflationary structural adjustment programs, usually under the auspices of the International Monetary Fund and the International Bank for Reconstruction and Develop-

ment (World Bank). At the same time, all of the countries in the region have had to come to terms with demands to alter radically their relations with the dominant industrial power in the area, South Africa.

The world's condemnation of apartheid and the brutally nondemocratic political system in South Africa has had profound effects not only on that country but also on social, political, and economic development throughout the region. Historically, the long-standing economic ties between the other countries in this region and South Africa made it difficult for those countries to attempt to sever relations with their powerful neighbour. The delegates to the first Southern African Development Coordination Conference (SADCC) in 1979 included Angola, Botswana, Mozambique, Tanzania, and Zambia. By 1980 Lesotho, Malawi, Swaziland, and the newly independent Zimbabwe had joined the group, which adopted the Lusaka Declaration at its first major regional economic summit in April of that year. An explicit aim of this declaration was to reduce the region's dependence on South Africa for transport, trade, and the supply of electric power.

Formation of SADCC

In 1986 SADCC adopted a resolution implementing economic sanctions against South Africa, but no timetable was prescribed. Moreover, several of the participating members actually increased their transport and trade links with South Africa during the late 1980s, while trade between SADCC members has remained much smaller than trade with South Africa.

The unbanning of the African National Congress (ANC) and the release of ANC leader Nelson Mandela in 1990 suggested the possibility of a resolution of major conflicts. There is now some optimism not only concerning future economic prospects within South Africa but also in the region as a whole.

Indicators of levels of development. Two widely used general development indicators—namely, the World Bank's ranking of economies from lowest to highest according to per capita income and the United Nations Development Program's ranking of countries according to the mortality of children under the age of five years—show that some of the countries in southern Africa are among the least developed in the world. South Africa and Botswana are clearly exceptions. In the early 1990s the World Bank ranked Mozambique as having the lowest per capita income in the world, while Malawi, Zambia, and Lesotho also fell into the category of "lowest income economies," with annual per capita incomes below $500 (U.S.). Angola, Swaziland, Zimbabwe, Namibia, and Botswana are categorized as "lower-middle-income economies," where income ranges between $600 and $2,400. South Africa just exceeds this level of per capita income and is the only country in the region classified as an "upper-middle-income economy." South Africa and Botswana have achieved much lower rates of under-age-five mortality than the other countries. Mozambique, Angola, and Malawi suffer from under-five mortality rates that are among the highest in the world.

These aggregate indicators of development must be treated with caution. The measurement of income is fraught with difficulties in economies where a high proportion of output is not marketed or accurately recorded, as is the case in all the economies in the region. Equally important, aggregate indicators shed no light on the crucial issue of the distribution of the benefits of growth. In South Africa, for example, the racially based historical pattern of development means that the living conditions of the majority of the population are poorly represented by the aggregate measures of income and child mortality. Nevertheless, it is important to stress that several countries in the region have achieved remarkable progress in the late 20th century with respect to some of the indicators of human welfare. All of these countries have seen a substantial improvement in life expectancy in the period since 1960. In Zimbabwe, Lesotho, and Botswana there have been dramatic changes in the number and proportion of children attending secondary schools, which will radically alter the skill composition of the labour force in the 1990s.

The regional gross national product (GNP) of the economies in southern Africa is dominated by the contribution of South Africa, which alone accounts for more than three-quarters of regional GNP; the next largest contribution to regional GNP is provided by Angola, followed by Zimbabwe and Zambia, each with only a few percent.

Population and land resources. The total estimated population for the region reached approximately 91 million by the end of the 1980s. South Africa had the largest population (in excess of 30 million), followed by Mozambique (15.3 million), Angola (9.7 million), and Zimbabwe (9.5 million). In contrast, the population was below 2 million in Namibia, Botswana, Lesotho, and Swaziland. Although rates of growth of population are generally high, they have also differed considerably between countries; these divergences are unlikely to narrow in the early 21st century.

The distribution of the population between rural and urban areas has altered dramatically in all of the countries in the region in the late 20th century. There has been a large increase in the share of the urban population. In South Africa and Zambia more than half of the population lives in urban areas, while in most of the other countries about 25 percent of the population is urban. In Malawi, which has the lowest proportion of total population living in urban areas, the share is 12 percent. Rates of urbanization have been high in all of the countries, with the urban population more than doubling between 1965 and 1990.

Increase in urban population

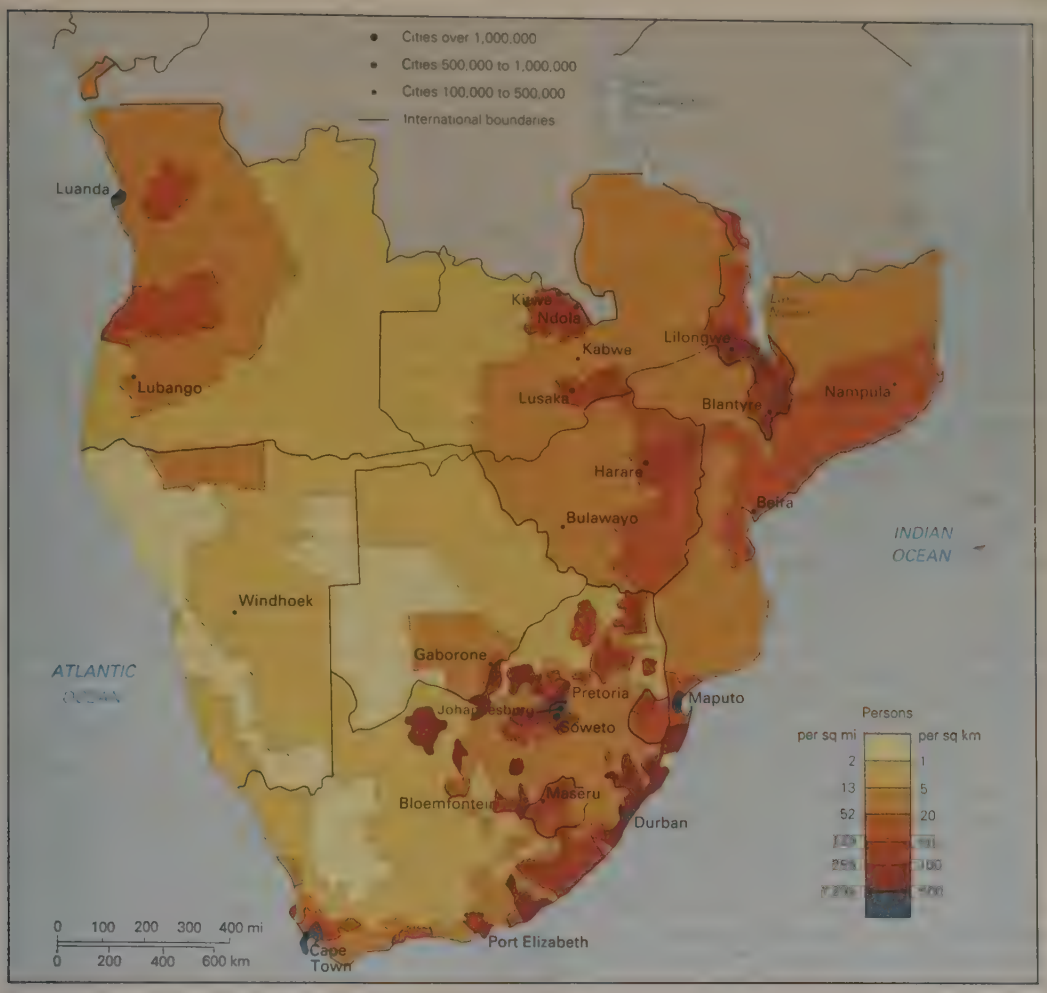
Angola and South Africa are the largest countries in the region in terms of land area; however, the proportion of each country's land that may be classified as arable differs considerably. The estimate for Angola is only 26 percent, compared to 77 percent for South Africa. In some countries the proportion of the potentially irrigable land that has been developed for irrigation is very low. For example, it is well below 10 percent in Angola, Zambia, Mozambique, and Malawi, suggesting considerable scope for increased agricultural production. In addition, these four economies, as well as Namibia, have substantial undeveloped fishing resources.

Attempts by the Food and Agriculture Organization of the United Nations to relate population density to agricultural production potential suggest that there are large differences in the degree of population pressure on land resources within the region. While Angola and Zambia clearly benefit from a large agricultural resource base relative to the size of their populations, Malawi, Lesotho, Botswana, Swaziland, and Namibia all show signs of acute pressure on available agricultural production resources. Most of the countries in the region were severely affected by drought in the early 1990s.

Many rural people can survive only by producing corn (maize), sorghum, cassava, and millet for their own consumption. A large proportion of those in the agricultural sector operate very small farms. Tenurial status on these farms and access to grazing land are often loosely defined and insecure. Yields are low and extremely variable, while access to improved inputs, such as fertilizers, pesticides, and seeds, and marketing facilities is limited. Nevertheless, throughout the region there is evidence of growing differentiation within the smallholder sector and of the growth of production for the market on the basis of wage labour as opposed to subsistence production and reliance on family labour.

In Zimbabwe, Swaziland, and South Africa much of the best land is monopolized by a few large estates, which also retain privileged access to inputs, especially credit. In Botswana, too, a freehold system dominated by large cattle ranches coexists alongside a sector characterized by small-scale producers and by forms of communal tenure. In Zambia a high proportion of marketed output and most of the land most favourably located for intensive agricultural production are controlled by a small number of farms. Although Mozambique and Angola have declared policies for the nationalization of all land and have invested in state farms as well as cooperative production structures, a high proportion of output is still produced within the smallholder sector. Privatization of state-run agricultural institutions has been an important feature of the structural adjustment and stabilization policies adopted throughout the region in the late 20th century. The success of these policies in terms of accelerated rates of growth of small-

Land holdings



Population density of southern Africa.

holder agricultural production has yet to be demonstrated, however.

The structure and growth of production. The conventional threefold division of production into agriculture, industry (including separate estimates for mining and manufacturing), and services shows that the majority of the countries in the region continue to remain reliant on the production of primary commodities. Almost 90 percent of production in Mozambique consists of raw materials, with approximately two-thirds being accounted for by the production of cotton, fish products, peanut (groundnut) oil, and sugar. Lesotho—which is predominantly a wheat producer—and Malawi—producing tea, tobacco, and sugar—also rely heavily on agriculture. Agriculture's share of the total output is high in Zambia (coffee, tobacco, and sugar) and Zimbabwe (tobacco, cotton, and sugar) as well, although in these two economies copper and other mining activities are of considerable importance. Botswana and Namibia depend to a great extent on the mining of diamonds, while South Africa (followed by Zimbabwe and Zambia) has the largest manufacturing sector in absolute and proportional terms.

Production of primary commodities

With regard to manufacturing, the only countries to have achieved substantial output by the beginning of the 1990s were South Africa, Zimbabwe, and, to a lesser extent, Zambia. While Zimbabwe's manufacturing concentrated on the processing of food, beverages, and tobacco, as well as textiles and iron ore, South Africa developed a significant manufacturing capacity in machinery and capital goods and in chemicals. The intensified threat of sanctions probably contributed to an acceleration in the growth of this capacity in South Africa in the 1970s and '80s. Not only does the manufacturing sector contribute far more to total output than is the case in the other economies in the region but the South African manufacturing sector is also much more integrated and diversified. Thus the contribution of machinery and transport equipment (20 percent) to South Africa's manufacturing value added is far larger than in other countries in the region. In Zimbabwe, with the second highest ranking, its share was below 10 percent at the beginning of the 1990s. The machinery and transport subsector barely existed in the other economies of the region.

Despite a diversified structure of production, the growth of output in the South African economy was rather low in the 1980s and early 1990s. The region as a whole has been characterized by widely diverging trends in the growth of output. Botswana has outstripped its regional neighbours, with a sustained average annual rate of growth of output in excess of 10 percent and has been one of the fastest-growing economies in the world. Lesotho and Zimbabwe have been growing by approximately 3 percent per year. South Africa has been growing at only about half this rate, and Zambia has grown even more slowly, at under 1 percent per year. Mozambique's output has declined during the same period. There does not appear to be a close relationship between the growth of output and the size of the public sector. Central government expenditure as a percentage of GNP varies widely within the region; however, in those economies that grew rapidly in the 1980s, central government expenditure accounted for a large proportion of GNP.

Major exports. Analysis of the region's major exports confirms the trends discussed above. All of the countries show a preponderance of raw materials in their export composition, in common with the majority of the economies of sub-Saharan Africa. There has been a period of decline in the demand for several types of metals and minerals from the major industrialized countries, partly as a result of the slowdown in the growth of these economies

and partly because of changes to production techniques that are less resource-intensive. Declines in demand have contributed to serious falls in the revenues from such commodities. With regard to agricultural exports, factors influencing their supply—such as unfavourable weather conditions, the persistence of wars, and the collapse of transport networks—have reduced foreign-exchange earnings. The existence of a rather diversified structure of agricultural exports, as in Mozambique, has not proved sufficient to prevent a decline of export revenues.

All the countries in southern Africa derive more than half their export earnings from raw materials. Angola, with its large deposits of petroleum, diamonds, and iron ore; Malaŵi, with its tobacco, tea, and sugar; and Zambia, with its copper, all obtain more than 90 percent of their export revenues from raw materials. South Africa and Zimbabwe derive lower proportions of their foreign-exchange earnings from the export of raw materials, reflecting the relatively developed manufacturing base in their economies and the contribution that manufactured exports make to their total exports.

Regional economic cooperation and growth. Despite SADCC's objective of reducing economic dependence on South Africa, the landlocked countries in the region still rely to a great extent on routes through South Africa for access to seaports and shipping. When SADCC was formed in 1980, approximately 80 percent of international commercial traffic from the six landlocked countries in the area went through South African ports. The evidence suggests that by 1990 approximately 60 percent of the transit from the landlocked member countries (Botswana, Lesotho, Malaŵi, Zambia, and Zimbabwe) was handled through SADCC ports. The ending of the wars fomented by the South African security forces in Angola, Mozambique, and Namibia should improve the viability of transport routes and reduce trading costs for Zambia, Botswana, Zimbabwe, and Lesotho. Following an internationally acceptable transfer of power in South Africa, the possibility of a cooperative transport agreement between SADCC members and South Africa will be a priority issue. Future trade prospects are likely to improve if the parties to such an agreement have unrestricted choices concerning transport routes.

International transportation

After the implementation of formal sanctions in the second half of the 1980s, trade with South Africa accounted for 30 percent of SADCC imports and 7 percent of exports. Zimbabwe, the dominant economy within SADCC, has remained particularly dependent on South Africa, trading far more with South Africa than with the SADCC countries. The attempts by SADCC to promote increased intra-member trade, which only amounted to 5 percent of total foreign-trade flows in the region by the end of the 1980s, may well receive less attention than the forging of new trade relationships with a democratic South Africa in the 1990s. In a new trading environment, it is possible that South Africa will play a positive role as a regional processor or beneficiation catalyst, helping to transform a range of raw-material exports into higher-value manufactured goods.

In such an environment, and with the prospects of more rapid economic growth, foreign investors would be far more likely to be positive with regard to the region as a whole. In addition, South African investors, who have considerable resources at their disposal, will be encouraged to turn their attention toward neighbouring economies. Increased integration, cooperation, and the reality of peaceful coexistence would eliminate the massive historic levels of expenditure on unproductive isolationist projects and the military. While many in the SADCC countries are now concerned about domination by the South African economy and the possibility of reductions in labour migration opportunities and the flow of remittances as a result of nationalist sentiment in a democratic South Africa, the prospect of rapidly expanding markets and capital inflows may serve to allay their pessimism.

Access to the South African labour market does, however, remain of critical importance to several economies in the region. For example, Lesotho, Mozambique, Botswana, and Swaziland were together still sending approximately a quarter of a million workers into South Africa at the beginning of the 1990s. The wages remitted by these workers are a major item in the balance of payments of many countries.

In addition, closer trading relationships with South Africa will require difficult policy decisions concerning the degree of protection to be afforded the small and insecure manufacturing sectors in neighbouring economies. Unrestricted access for manufactured imports could result in the disappearance of recently established manufacturing capacity that might, in the longer run, have proved viable and competitive. More positively, during the 1990s South Africa's neighbours will have enhanced access to a far more sophisticated industrial infrastructure and more advanced communications and information systems, as well as to African-based skilled personnel with expertise in the adaptation of agricultural and industrial technologies to the southern African environment. (J.B.S./H.Za.)

History

The history of southern Africa cannot be written as a single narrative. Shifting geographic and political boundaries and changing historiographical perspectives render this impossible. Research into local history in the last decades of the 20th century fragmented historical knowledge, and older generalizations gave way to a complex polyphony of voices, as new subfields of history—of gender and sexuality, health, and the environment to name but a few—developed. Divided societies produce divided histories, and there is hardly an episode in the region's history that is not now open to debate. In southern Africa, history is not a set of neutrally observed and agreed-upon facts: present concerns colour interpretations of even the remote past. For all the contestants in contemporary southern Africa there has been a conscious struggle to control the past in order to legitimate the present and lay claim to the future. Who is telling what history for which Africa is a question that needs constantly to be addressed.

EARLY HUMANS AND THE STONE AGE

The controversies in southern African history begin with the discovery of a fossilized hominid skull in a limestone cave at Taung near the Harts River north of Kimberley in 1924, followed from 1936 by discoveries in similar caves in the Transvaal and northern Cape. For some time the significance of these finds and their relationship to the evolution of early humans was unappreciated, perhaps because the finds could not be dated, and stone tools—long regarded the defining characteristic of early humans—had not been found with them. Since that time, similar, more datable discoveries in eastern Africa have made it possible to place the South African remains in sequence and identify them as australopithecines, upright-walking creatures who are the earliest human ancestors. The australopithecines who roamed the highland savanna plains of southern Africa date from about three to one million years ago. There can be little doubt that for hundreds of thousands of years southern Africa, like eastern Africa, was in the forefront of human development and technological innovation.

Controversies remain, however. The connections between australopithecines and earlier potentially hominid forms remain unclear, while a number of different species of australopithecine have been identified. Their evolution into the species *Homo habilis* and then into the species *Homo erectus*—which displayed the larger brain, upright posture, teeth, and hands resembling those of modern humans and from whom *Homo sapiens* almost certainly evolved—is still fiercely debated. *Homo erectus* appears to have roamed the open savanna lands of southern and eastern Africa, collecting fruits and berries—and perhaps roots—and either scavenging or hunting. Paleolithic Period (Early Stone Age) technology—known as Acheulian industry and characterized by the use of simple stone hand axes, choppers, and cleavers—seems to have spread from eastern Africa throughout the continent and also to Europe and Asia during the Middle Pleistocene Epoch, reaching southern Africa about half a million years after its first

Homo erectus

appearance in eastern Africa about 1.5 million years ago. Acheulian industry remained dominant for more than one million years.

During this time early humans also developed those social, cognitive, and linguistic traits that distinguish *Homo sapiens*. Some of the earliest fossils associated with *Homo sapiens*, dated from about 120,000 to 80,000 years ago, have been found in South Africa at the Klasies River Mouth Cave in the eastern Cape, while at Border Cave on the South Africa–Swaziland border a date of about 90,000 years ago has been claimed for similar Mesolithic Period (Middle Stone Age) skeletal remains.

With the emergence of *Homo sapiens*, experimentation and regional diversification displaced the undifferentiated Acheulian tool kit, and a far more efficient small blade, or microlithic, technology evolved. Through the controlled use of fire, denser, more mobile populations could move for the first time into heavily wooded areas and caves. Wood, bark, and leather were used for tools and clothing, while vegetable foods were also probably more important than their archaeological survival suggests.

Hunter-gatherers

Some scholars believe that the addition of organized hunting to gathering and scavenging transformed human society. The large number of distinctive Later Stone Age industries that emerged reflect increasing specialization as hunter-gatherers exploited different environments, often moving seasonally between them, and developed different subsistence strategies. As in many parts of the world, changes in technology seem to mark a shift to the consumption of smaller game, fish, invertebrates, and plants. Later Stone Age peoples used bows and arrows and a variety of snares and traps for hunting, as well as grindstones and digging sticks for gathering plant food; with hooks, barbed spears, and wicker baskets they also were able to catch fish and thus exploit rivers, lakeshores, and seacoasts more effectively.

Despite the ever-increasing number of radiocarbon dates available for the many Later Stone Age sites excavated in southern Africa, the reasons for changed consumption patterns and variations in technology are poorly understood. Nevertheless, the appearance of cave art, careful burials, and ostrich eggshell beads for adornments suggests more sophisticated behaviour and new patterns of culture. These developments apparently are associated with the emergence between 20,000 and 15,000 BC of the earliest of the historically recognizable populations of southern Africa: the Pygmy, San, and Khoi peoples.

Although many scholars attempt to deduce the nature of Later Stone Age societies by examining contemporary hunter-gatherer societies, this method is fraught with difficulties. Evidence from Botswana and Namibia suggests that many contemporary hunter-gatherers recently have been dispossessed, and that their present way of life, far from being the result of thousands of years of stagnation and isolation, has resulted from their integration into the modern world economy; this hardly provides an adequate model for reconstructions of earlier societies.

During historic times hunter-gatherers were organized in loosely knit bands, of which the family was the basic unit, although wider alliances with neighbouring bands were essential for survival. Each group had its own territory, in which special importance was attached to natural resources, and in many instances bands moved seasonally from small to large camping sites, following water, game, and vegetation. Labour was allocated by gender, with men responsible for hunting game, women for snaring small animals, collecting plant foods, and undertaking domestic chores. These patterns are also evident in the recent archaeological record, but it is unclear how far they can be safely projected back.

Contrary to the popular view that the hunter-gatherer way of life was impoverished and brutish, Later Stone Age people were highly skilled and had a good deal of leisure and a rich spiritual life, as their cave paintings and rock engravings show. Dating is problematic, but at Apollo 11 Cave in southern Namibia paintings have been dated to some 28,000 years ago. Whereas the art in the northern woodlands is stylized and schematic, that of the savanna and coastlands seems more naturalistic, showing scenes of hunting and fishing, of ritual and celebration. The motives of the artists remain obscure, but many paintings appear linked to the trance experiences of medicine men, in which the eland was a key symbol. In later rock paintings there is also the first hint of the advent of new groups of herders and farmers.

FOOD PRODUCTION

In the long run these new groups of herders and farmers transformed the hunter-gatherer way of life. Initially, however, distinctions between early pastoralists, farmers, and hunter-gatherers were not overwhelming, and in many areas they coexisted. The first evidence of pastoralism in the subcontinent occurs in the more arid west; there the bones of sheep and goats, accompanied by stone tools and pottery, date to some 2,000 years ago, about 200 years before iron-using farmers first arrived in the better-watered eastern half of the region. It is with the origins of these food-producing communities and their evolution into contemporary societies that much of the precolonial history of the subcontinent has been concerned.

When Europeans first rounded the Cape of Good Hope, they encountered herding people, whom they derogatorily called "Hottentots" but who called themselves "Khoikhoin" meaning "Men of Men." At that time Khoikhoi inhabited the fertile southwestern Cape region as well as its more arid hinterland to the northwest, where rainfall did not permit crop cultivation, but they may once have grazed their stock on the more luxuriant central grasslands of southern Africa. Linguistic evidence suggests that the languages of the later Khoikhoi (the so-called Khoe languages) originated in one of the hunter-gatherer languages of northern Botswana. In the colonial period, destitute Khoikhoi often reverted to a hunter-gatherer existence; herders and hunters were also frequently physically indistinguishable and used identical stone tools. Thus, the Dutch, and many subsequent social scientists, believed they belonged to a single population following different modes of subsistence: hunting, foraging, beachcombing, and herding. For this reason the groups are often referred to as Khoisan, a compound word referring to Khoikhoi and San, as the Khoikhoi called hunter-gatherers without livestock ("Bushmen" in the terminology of the colonists).

Khoikhoi

Most of southern Africa's early agricultural communities shared a common culture, which spread across the region remarkably quickly from the 2nd century AD. By the second half of the 1st millennium AD, farming communities were living in relatively large, semipermanent villages. They cultivated sorghum, millet, and legumes and herded sheep, goats, and some cattle; made pottery and fashioned iron tools to turn the soil and cut their crops; and engaged in long-distance trade. Salt, iron implements, pottery, and possibly copper ornaments passed from hand to hand and were traded widely. Some communities settled near exceptionally good salt, metal, or clay deposits or became known for their specialist craftsmen.

Archaeologists are divided over whether all these cultural and economic attributes arrived with a single group of new immigrants speaking a new language or resulted from a more piecemeal development of different skills and the adoption of new techniques by indigenous hunter-gatherers, as has already been suggested in the case of herding among the Khoikhoi. Moreover, archaeologists disagree about the routes and modes of dispersal as well as its timing. It seems likely, however, that a movement of immigrants into southern Africa occurred in two streams and was part of a wider expansion of populations speaking Bantu languages that ultimately derived from the Niger-Congo languages of western Africa.

Bantu speakers

"Eastern-stream" Bantu speakers, associated with the earliest farming communities in the well-watered eastern half of southern Africa, date from the 2nd to 5th century AD. Similar pottery has been found stretching from north-eastern Tanzania and coastal Kenya into eastern South Africa, Mozambique, and Swaziland. These early farmers settled on arable soils along coastal dunes, rivers, and valley basins. Where possible, they exploited marine resources, planted cereals, and worked iron, although cattle and long-distance trade were insignificant.

"Western-stream" Bantu speakers were initially more familiar with fishing, oil palms, and the cultivation of vegetables than with cereals or cattle. Even before the 1st millennium AD, pottery similar to that of the Eastern stream was being made in the upper Zambezi valley, and pottery of a slightly more recent date has been found in parts of northern Angola. It was probably from these communities that the Bantu speakers spread into the more arid western half of the subcontinent, northwestern Zambia, southwestern Zimbabwe, along the eastern margins of the Kalahari Basin into Botswana, and later into eastern South Africa and Mozambique. Like their counterparts in the east, western-stream Bantu speakers cultivated cereals, worked metal, and made pottery, but the evidence of livestock is far more clear-cut; at first they primarily raised sheep and goats, slightly later cattle.

Although at first the impact of food production was probably less momentous than is often assumed, agriculture combined with pastoralism and metallurgy could support far larger settled communities than previously had been possible and enabled a more complex social and political organization to develop. Cattle raising led to increased social stratification between rich and poor and established new divisions of labour between men and women; the accumulation of cattle and the continuous site occupation inherent in cereal production enabled the storage of wealth and the deployment of more organized political power. The relationships established among hunters, herders, and agriculturalists over more than 2,000 years of socioeconomic change ranged from total resistance to total assimilation. For the indigenous people of southern Africa the frontiers between different modes of subsistence presented new dangers and opportunities.

Social stratification

As the new culture spread, larger, more successful farming communities were established; in many areas the new way of life was adopted by the hunter-gatherers. Even in the apparently inhospitable and isolated Kalahari it is now clear that there was intense interaction and exchange between hunter-gatherers and food-producers, leading to the development of hybrid amalgams of pastoralism, agriculture, and foraging. Contemporary Bantu-speaking peoples of southern Africa are genetically very similar to the Later Stone Age people of Africa; their close relationship also is evidenced by the presence of Khoisan "click" sounds and loanwords in southeastern Bantu.

THE RISE OF MORE COMPLEX STATES

From about the turn of the 1st millennium AD, in some areas of what are now central Zambia, southeastern Zimbabwe, Malaŵi, and eastern South Africa, changes in ceramic style were paralleled by a change in the location and nature of settlements. More sophisticated techniques of ironworking, more extensive gold and copper mining, and a great increase in stone building suggest the evolution of more complex state structures, the growth of social inequalities, and the emergence of new religious and spiritual ideas. These changes were, however, neither simultaneous nor evenly spread.

The nature of these transitions and the differences among the sites are still poorly understood, and, again, archaeologists disagree as to whether the changes can be explained by local developments or by the arrival of migrating populations. In part the controversy may reflect regional differences. In most of Zambia and Malaŵi a sharply distinguishable pottery style appears at this time, probably from southeastern Zaire, and forms the basis of the ceramics made by several different societies. Farther west, however, there are greater continuities with the earlier wares, while in southeastern Africa locally driven increases in population and cattle—which led to expansion into less favourable environments but which also brought new ideas and new methods of political control—may hold the key.

Toutswe

Whatever the explanation, many of the changes appear for the first time at Toutswe in eastern Botswana with the appearance about the 7th century AD of a new ceramic tradition, new technology, and new forms of social and economic organization. There, larger, well-defended hilltop capitals probably dominated a series of smaller sites with access to water over a wide region. Toutswe may betoken the advent of a new population; on the other hand, the evidence of its large cattle herds provides insight into the way in which the natural buildup of herds in a favourable environment could stimulate social change and territorial expansion. Cattle underpinned both material and symbolic power in southern Africa and served to cement social obligations through bridewealth and loan arrangements. Cattle were also an ideal medium for exchange, and the increase in herding necessitated increased specialization and the extension of trading networks. Patrilineal and polygynous cattle-keeping farmers thus had immense advantages over communities that lacked these new forms of wealth and social organization. Similarities between Toutswe and the material culture of later sites in the Limpopo valley and Zimbabwe suggest that Toutswe also may have inspired new forms of social and economic organization for peoples further afield.

Greater stratification and more complex social organization were also probably accelerated by the growth of trading with the outside world and by competition for access to it. In the early centuries AD the northeastern African coast was well known to the traders of the Greco-Roman world. These contacts diminished with the rise of Islām, and the east coast became part of the Indian Ocean trading network. By the 8th century Arab traders had begun to visit more southerly harbours, and between the 11th and 15th centuries they founded some three dozen new towns. Although they never united politically, these towns developed a common Afro-Arabic, or Swahili, culture and a splendour that amazed the first European arrivals.

The Limpopo and Save rivers were early arteries of the trade from the southernmost Arab trading posts, with African intermediaries initially bringing ivory and perhaps animal skins, and later copper and gold, to the coast. Persian potsherds at Chibuene on the Mozambique coast in the 8th century and the presence of snapped cane glass beads as far in the interior as the Kruger National Park and at Schroda on the Limpopo, as well as in Botswana, on the Zimbabwe plateau, and as far south as the Mngeni River near Durban, all attest to the influence of this long-distance trade in the region and its early integration into the Indian Ocean networks.

At 9th- and 10th-century sites such as Schroda and Bambandyanalo in the Limpopo valley, the ivory and cattle trade seems to have been of major importance, but later Iron Age sites such as Mapungubwe (a hilltop above Bambandyanalo), Manekweni (in southwestern Mozambique), and Great Zimbabwe, which date from the late 11th to the mid-15th century, owed their prosperity to the export of gold. Farther north, the 14th-century site of Ingombe Ilede (near the Zambezi-Kafue confluence) probably also owed its prosperity in copper and gold—and its social stratification—to the rise of the east coast trade. Although they do not typify the later Iron Age as a whole, the conspicuous consumption at these sites and the bias in oral sources toward centralized states means they have attracted perhaps a disproportionate share of scholarly attention.

Mapungubwe, Manekweni, and Great Zimbabwe

In Mapungubwe and Great Zimbabwe a wealthy and privileged elite built with stone and were buried with gold and copper ornaments, exotic beads, and fine imported pottery and cloth. Their homes, diet, and ostentatious burials are in stark contrast to those of the common folk, whose dwellings cluster at the foot of the sites where they probably laboured. Large quantities of stone were brought to build walls on these hilltop sites, which suggests considerable labour. All were centres of political authority, controlling trade and cattle movement over a wide area stretching from eastern Botswana in the west to Mozambique in the east. Cattle, gold, and copper came in trade or tribute from settlements hundreds of miles away. Skilled craftsmen made elegant pottery, sculpture, and fine bone tools for local use and for trade, while the presence of spindle whorls suggests local weaving.

In the past, fierce controversy raged concerning the racial identity of Mapungubwe's occupants, and, as in the case of Great Zimbabwe, early excavators refused to accept that it could have been built by Africans. Mapungubwe's skeletal and cultural remains are, however, identical to those found at other Iron Age settlements in the subconti-

nent, and there is little reason to doubt the African origin and medieval date of both sites.

In the second half of the 15th century Great Zimbabwe came to an abrupt end. Its successor in the southwest was Torwa, with its centre at Khami; in the north it was replaced by the Mutapa state. The new culture at Khami developed both the stone-building techniques and the pottery styles found at Great Zimbabwe and seeded a number of smaller sites over a wide region of the southern and western plateau. The Torwa kingdom seems to have lasted until the end of the 17th century, when it was replaced by the Rozwi Changamire dynasty from the central plateau, which lasted well into the 19th century. The domination of the Mutapa state extended into Mozambique. Like the rulers of Great Zimbabwe, the Torwa, Mutapa, and Rozwi dynasties maintained the coastal gold and ivory trade, although cereal agriculture and cattle remained the basis of the economy.

In the first half of the 2nd millennium AD the majority of southern Africa's peoples were probably relatively unaffected by the formation of these larger trading states. Most lived in small-scale societies, based on kinship, in which political authority was exercised by a chief who claimed seniority by virtue of his royal genealogy but who may have risen to power through his access to mineral resources, hunting, or ritual skills. By 1500 most of the farming communities had stabilized in roughly their present-day habitats, reaching their ecological frontier on the dry southern Highveld of South Africa and gradually clearing the coastal forests.

Population growth

While in many areas ceramic evidence suggests cultural continuity over many centuries, within these boundaries there was considerable movement as populations expanded and found available resources inadequate. Thus, between the 17th and 19th centuries there was migration of northern and eastern Shona speakers into the centre and south of the plateau, while in South Africa new land was colonized by cattle-raising peoples, as the stone-walled sites in the southern Highveld indicate. In some areas the expansion inevitably led to conflict as the newcomers came up against settled communities; in others the indigenous inhabitants were gradually absorbed, while elsewhere sparsely inhabited, colder, and more arid mountain lands were colonized.

In most of these farming communities land was relatively plentiful, while labour was not, and control over people was therefore of the essence. Those societies in which cattle were important were patrilineal, polygynous, and virilocal; men herded, while women were the major agricultural producers. The labour and reproductive power of women was transferred from father to husband through the circulation of cattle in the form of bridewealth. Where cattle were meagre, societies were matrilineal and usually matrilocal; men still depended on women for agricultural labour and for bringing young men and children into the household. Wealthy homes were those with large numbers of women, and even before the advent of the Atlantic slave trade it had become customary for men to take slave wives who would work in exchange for protection.

By the time coastal peoples were first encountered by literate European observers in the 15th century, many were the recognizable forebears of southern Africa's contemporary population. This does not mean, however, that these societies were static and unchanging. New kingdoms and chiefdoms were formed and older ones disintegrated, the result of both internal and external agency, while new ethnic and cultural identities began to be forged in the hazardous new world resulting from Africa's incorporation into the Atlantic economy.

THE COMING OF THE PORTUGUESE

The first of these literate newcomers were the Portuguese, who from the 15th century edged their way around the African coast in the hope of outflanking Islām, finding a sea route to the riches of India, and discovering additional sources of food. They reached the kingdom of the Kongo in northwestern Angola in 1482–83; early in 1488 Bartolomeu Dias rounded the southern tip of the continent; and just over a decade later Vasco da Gama sailed along the east coast of Africa before striking out to India. Unpropitious as they were initially, these voyages marked the beginning of the integration of the subcontinent into the new world economy and the dominance of Europeans over the indigenous inhabitants.

The Portuguese in west-central Africa. Portuguese influence in west-central Africa radiated over a far wider area and was much more dramatic and destructive there than on the east coast. Initially the Portuguese crown and Jesuit missionaries forged peaceful links with the kingdom of the Kongo, converting its king to Christianity. Almost immediately, however, slave traders followed in the wake of priests and teachers, and west-central Africa became tied to the demands of the São Tomé sugar planters and the transatlantic slave trade.

The slave trade

Until 1560 the Kongo kings had an effective monopoly in west-central Africa over trade with metropolitan Portugal, which showed relatively little interest in its African possessions. By the 1520s, however, Afro-Portuguese traders and landowners from São Tomé were intervening in the affairs of the Ndongo kingdom to the south, supporting the ruler, or *ngola,* in his military campaigns and taking his war captives and surplus dependents as slaves. By the mid-16th century Ndongo, with Portuguese assistance, had become a major kingdom extending over a wide area between the Dande, Lukala, and Kwanza rivers.

By the last third of the 16th century, the Portuguese attitude toward Africa had changed; rumours of fabulous gold and silver to be found in the interior led in 1569 to the dispatch in the east of Francisco Barreto to discover the sources of gold in the Mutapa kingdom and to the appointment in 1575 of Paulo Dias de Novais to search for what turned out to be mythical silver mines in the west. Dias established himself as captain-general, or governor, in Luanda, with jurisdiction over an undefined area between the Dande and Kwanza rivers. Within a few years of his arrival the first war of a century of almost constant warfare had begun. The wars soon resolved themselves into slave-raiding campaigns, as Europeans demanded labour rather than tropical products in exchange for their merchandise, and African societies rapidly exhausted available supplies of war captives and criminals.

Chiefs exchanged slaves for European firearms and luxury goods and secured further dependents with cheaply produced textiles and fiery Brazilian alcohol. Impelled by the increased demand for slaves for the sugar plantations of São Tomé and later Brazil, and relying on African mercenaries and allies, the military governors of Luanda launched armed incursions against the people of the interior. Kingdoms rose and fell as African rulers were ineluctably drawn into the slave trade and were as often destroyed by it.

New warlords emerged at the head of bands of starving refugees, who from the late 16th until the 18th century swarmed down from the hills, fought one another, and devastated the settled kingdoms. By the end of the 16th century, well-organized military bands of marauders, known as the Imbangala, began to appear along the coast south of Luanda. In their anxiety to swell slave numbers, Portuguese governors allied with these war bands, and together they dealt the final blow to the Ndongo kingdom about 1622. By that time the Imbangala had retreated to the middle Kwango, where they founded the kingdom of Kasanje. Over the next two centuries this kingdom replaced Ndongo as the chief slave-trading entrepôt between the coast and the east, where the highly centralized and militarist Lunda kingdoms became increasingly important in supplying slaves by the 18th century.

As the Portuguese were penetrating inland from Luanda at the beginning of the 17th century, they also moved southward. In 1617 they established a colony at Benguela, which, as in the case of the Kongo kingdom, was annexed as part of Angola in the 19th century. Expansion inland from Benguela, however, like the initial expansion farther north, was spearheaded by Afro-Portuguese slave traders, who used southern ports to outflank Portuguese control. As the slave frontier moved south, so the process of constructing and then destroying slave-trading warrior kingdoms was repeated. Those who were not crushed by the process

sought safety in woodlands and swamps or joined new heterogeneous communities of refugees, like the Chokwe ("Those Who Fled") of the western savanna. These new communities often became slave-raiders themselves.

Through the 18th and early 19th centuries the slave trade remained at the centre of Angola's economic existence, with Benguela replacing Luanda as the chief port. As a result, the Ovimbundu kingdoms on the Bié Plateau, which probably were formed by refugees from the Imbangala and Mbundu kingdoms in the late 16th and 17th centuries, displaced Kasanje as the main source of slaves. The expansion of plantations in the New World doubled the numbers of slaves exported in the last third of the 18th century, when trade routes stretched as far as the Kunene River in the south and met up with the routes from Mozambique in the heart of Africa.

Impact of the slave trade

It is not possible to compile an exact balance sheet of the devastation caused to west-central Africa by the slave trade, and historians differ in their estimation of the numbers involved and of the extent of the damage inflicted. In the 17th century some 10,000 to 12,000 slaves were exported annually from Luanda. Although this figure includes captives from both north and south of the bay, it does not include those smuggled out to escape official taxation. In the 18th century about a third of the slaves exported to the Americas probably came from Angola. The figure probably represents a relatively small proportion of the total population of a huge area in any one year, but it was a significant proportion of economically active adults. The figure also does not take account of the depopulation and social dislocation resulting from incessant warfare and banditry, resulting famine and disease, and the intensification of slavery within African society.

The better-watered regions may have recouped their population losses within a couple of generations, supported by the introduction of new food crops such as manioc and corn (maize), which the Portuguese imported from South America. Nevertheless, the effects of the slave trade were, in social terms, incalculable. Accounts of Ndongo as rich and populous in the 16th century gave way to lamentations about its desolation in the 17th. The processes of border raids, wars of conquest, and civil strife, which affected the Ndongo and then the kingdoms of the Kwango River valley in the 17th century, were repeated to the south and east in the course of the 18th century as the slave frontier expanded. The ending of more overt violence as the slave frontier moved on left the weak—women, children, and the poor—vulnerable to innumerable personal acts of kidnapping and betrayal, a process exacerbated by the indebtedness of local traders to coastal merchants and the dependence of the traders on the transatlantic economy.

The Portuguese in southeastern Africa. Initially the southeastern coast was of far less concern to the Portuguese than west-central Africa. Within a few years of their arrival, however, they had seized its wealthy but divided cities and had established themselves on Moçambique and Sofala, which soon became key ports of call for ships on the way to India.

The Portuguese conquests led to the economic and cultural decline of the east coast cities. Yet the newcomers soon discovered that they were unable to control the vast area they had conquered. They faced resistance from coastal communities throughout the 16th century, and the profits they expected from the gold trade failed to materialize. In an attempt to control the trade and to discover the precious minerals for themselves, the Portuguese, following in the tracks of Muslim traders from the coast, expanded into the Zambezi valley about 1530.

Expansion into the Zambezi valley

In the Zambezi valley the Portuguese penetrated the Mutapa kingdom, with its heartland in the northeast between the Zambezi and Mazoe rivers. By the 1530s the Portuguese dominated the trade exits from the coast and had established fortresses and trade fairs along the Zambezi and on the plateau, where Africans came to exchange ivory and gold for beads and cloth. After 1541 Portuguese residents at these outposts elected representatives who were delegated certain powers by the Mwene (ruler of) Mutapa. Individual Portuguese and Goans also were able to get land grants and judicial rights from local rulers, which enabled them to extract tribute from the local population. These early grants formed the basis of what became known as the *prazo* system of landholding. Between the 17th and 19th centuries *prazeros* became immensely powerful and interfered in local African politics, creating an Afro-Portuguese society in the lower Zambezi valley independent of either African or Portuguese jurisdiction. Assisted by slave-soldiers known as the Chikunda, Afro-Portuguese warlords engaged in the slave and ivory trade, unsettling a wide area of east-central Africa.

The effect of Portuguese traders along the Zambezi valley on the Mutapa state was minimal until the late 16th century. In the 1560s, however, their hold was probably strengthened with the appearance in Zambesia of people known as the Zimba, a term applied to any marauders. These seem to have been Maravi people, who had first migrated from Luba territory to the southern end of Lake Nyasa (Lake Malaŵi) in the 14th century. There they broke up into a number of chiefdoms, usually under the paramountcy of the most powerful chief, who controlled the rain shrine at the heart of the local religion. The reasons for the Zimba irruption are far from clear, however. The Maravi attacked chiefs friendly to the Portuguese, as well as their settlements at Sena and Tete and on the coast. By 1601 the Mwene Mutapa was forced to call on the Portuguese for assistance, and this led to almost a century of increasingly disruptive Portuguese intervention in the affairs of Shona kingdoms to the south of the Zambezi.

Although attempts to drive the Portuguese from the Zambezi were unsuccessful until the late 17th century, when they were driven out by the armies of the Rozwi kingdom, this appearance of Portuguese power was deceptive: the Portuguese never had the resources to control the interior, and it was the Afro-Portuguese *prazeros* and the Rozwi Changamire dynasty who truly exploited the Mwene Mutapa's weakness.

In addition to gold, the Portuguese were interested in ivory and other mineral resources of the eastern African interior, particularly after 1700, when the gold appeared exhausted. A search for silver mines had first led them into Malaŵi in the 17th century, and from that point there is direct, though fragmentary, evidence of developments in the region. While the Portuguese records suggest that before 1590 there were no large states in the region, by the first decades of the 17th century a powerful kingdom had emerged under Muzura, perhaps out of an earlier system of small Maravi states at the southern end of Lake Nyasa.

By mid-century Muzura was eclipsed by the Kalonga, whose capital lay on the southwestern shore of Lake Nyasa, while by the turn of the 18th century the rise of the well-armed Yao in the trade between Lake Nyasa and the coast, and of the Bisa as middlemen to the west, contributed to the disintegration of the Maravi confederacy into several more or less autonomous fragments. This process was further accelerated by the wars and slave raids of the 19th century and the advent of the missionaries. By the early 18th century the Portuguese also had penetrated into present-day Zambia, establishing trading fairs at Zumbo and Feira on the Zambezi. Although there were no highly organized broker kingdoms in the area, *prazeros* traded gold and slaves to the coast.

Dutch and British competition

As in west-central Africa, from the beginning of the 17th century the Portuguese faced increasingly severe competition from Dutch and British ships in the Indian Ocean, while north of Cape Delgado the Arabs also took advantage of Portuguese weakness. In 1631 a series of revolts began on the east coast; by the beginning of the 18th century the Portuguese had been driven from the coast north of the Rovuma River. The Portuguese then turned their attention southward, where they had traded at Delagoa Bay with the local Tsonga inhabitants since the mid-16th century. They were unable to establish themselves at the bay permanently, however, and through the 18th century Dutch, English, and Austrian ships competed for the local ivory while North American whalers also traded there for food and cattle. Local chiefdoms vied for this market, and this competition contributed to the buildup of larger states in the hinterland of Delagoa Bay from the mid-18th century. Doubtless there was also trade in slaves, although

the numbers seem to have remained relatively small before the 19th century.

THE DUTCH AT THE CAPE

Apart from the Portuguese enclaves in Angola and Mozambique, the only other area of European settlement in southern Africa in the 17th and 18th centuries was the Dutch settlement at the Cape of Good Hope. In the late 16th century the Cape had become a regular port of call for the crews of European ships, who found local people ready to barter cattle in exchange for iron, copper, beads, tobacco, and brandy. By the mid-17th century Khoi intermediaries traded far into the interior. These trade relationships profoundly affected the nature of contact between the Khoikhoi and the Dutch.

In 1652 the Dutch East India Company dispatched Commander Jan van Riebeeck and 125 men to set up a refreshment station at the Cape. This outpost soon grew into a colony of settlement. In 1657 the company released a number of its servants as free burghers (citizens) in order to cultivate land and herd cattle on its behalf, and in the next year the first shiploads of slaves arrived in the colony. Although the company prohibited the enslavement of the local inhabitants, in order to protect the cattle trade, the loosely organized Khoikhoin were soon undermined by the incessant Dutch demands for their cattle and encroachment on their grazing lands and waterholes. The climate of the Cape was well suited to Europeans and their birth rate was high; whereas in Angola and Mozambique the Portuguese were ravaged by disease, at the Cape it was the indigenes who were decimated by the smallpox, influenza, and measles epidemics brought by Europeans.

Dutch expansion. By the end of the 18th century, Cape settlers—called Boers (Dutch *boer,* "farmer")—were far more numerous than their Lusophone counterparts, owing largely to natural increase. Men outnumbered women by three to two. Despite the varied European origins of the settlers, their shared vicissitudes and the company's insistence that all settlers speak Dutch and practice Calvinism led to a certain cultural uniformity and sense of group identity. The settlers began to call themselves "Afrikaners"—Africans. Nevertheless, class divisions in Cape Town and its environs were marked. A small group of affluent merchants and status-conscious company servants lived in Cape Town; in the neighbouring farming districts of the southwestern Cape, a wealthy gentry used slave labour to produce wine and wheat for passing ships. Independent small farmers eked out a living on the land, and a number of landless whites worked for others, generally as supervisors.

Trekboers

In the arid interior, economic necessity and ecology dictated a pastoral way of life for the Dutch cattle farmers, or trekboers. The poor soil and inadequate rainfall of the region necessitated vast, scattered farms, and the white population was thus thinly spread over an immense area. Although earlier literature stresses their mobility and subsistence economy, most frontier families occupied the same farms during their lifetime and remained dependent on the market for essentials such as arms and ammunition as well as for luxuries like tea, coffee, tobacco, and sugar.

The greatest barrier to Dutch expansion was the range of mountains inland from Cape Town. Once these were crossed and Khoisan resistance overcome, trekboers expanded rapidly to the east and north. The company made only sporadic attempts to follow them. Governmental authority was weak, and on the frontier trekboers were left to crush Khoisan resistance and mount their own defense through the commando system. They became accustomed to ruling over their slaves and Khoisan servants and clients as they saw fit, often with a ferocity born of fear. As the settlers expanded, their impact—through forced trade, plunder, and human and cattle disease—was increasingly destructive for the inland Khoisan, who retaliated by stealing settler cattle and burning homesteads.

The number of slaves increased along with the settler population, especially in the arable districts. Experiments in the use of indentured European labour were unsuccessful, and by the mid-18th century about half the burghers at the Cape owned at least one slave, though few owned more than 10. Slaves spoke the creolized Dutch that in the 19th century became Afrikaans. Many adopted Islām, which alarmed the ruling class. Divided in origin and dispersed geographically, slaves did not establish a cohesive culture or mount effective rebellions. Individual acts of defiance were frequent, however, and in the early 19th century there were two small uprisings. Nevertheless, in Cape Town itself slave culture provided the basis for a working-class culture after emancipation.

Slavery at the Cape is often portrayed as benign, but mortality rates were high and birth rates low; punishments for even minor misdemeanors were fierce, perhaps because adult male slaves greatly outnumbered their owners. Manumission, baptism, and intermarriage rates were also low, although newcomers and poorer burghers married slave women and, more rarely, Khoi women. Cohabitation with indigenous women was more common, especially in frontier districts where there were few white women. The children of these interracial unions, however, took on the unprivileged status of their mothers, so the practice did not affect the racially defined class structure of the society forming at the Cape. By the late 18th century in the Cape most blacks were "servants" and most Europeans were "masters."

The existence of slavery affected the status and opportunities of the dispossessed Khoisan who entered the labour market in increasing numbers from the late 17th century. Although theoretically they were free, compulsion governed the relationship between master and servant, and the legal status of the Khoisan increasingly approximated that of slaves. As the Cape became increasingly involved in the world economy, the demand for food for European ships escalated, as did calls for increased controls over Khoisan labour: in 1775 a system of "apprenticing" Khoisan children until the age of 25 was established, and by the end of the century the Khoisan were subject to a pass system similar to that which curtailed slave mobility. As they lost their cattle and grazing areas, the Khoikhoi became virtual serfs on settler farms, although some groups managed to escape beyond colonial borders.

Khoisan resistance

Khoisan resistance to Dutch colonialism erupted into guerrilla war on three occasions in the 17th century; the first, in 1659, nearly destroyed the settlement. Cattle raids punctuated almost every decade of the 18th century. The raids and counterraids became increasingly violent as the Dutch expanded into the fine sheep country to the northeast. Between 1799 and 1803 dispossessed Khoisan farmworkers in Graaff-Reinet, many with horses and guns, rose in revolt, challenging the entire colonial order. The Dutch feared that the Khoisan would attack the arable farms of the southwest, especially as they were joined by Xhosa allies. The intervention of government troops, divisions among Khoisan and Xhosa forces, and sheer bloodletting led to the defeat of the uprising, although it haunted the colonial imagination well into the 19th century. This was the last time the Khoisan fought under their traditional leaders to regain their lost lands.

The Cape's eastern frontier. For more than a century, the eastern frontier of the Cape was at the heart of the colonial encounter. Settler expansion to the east was blocked when trekboers came up against populous Xhosa farmers in the area of the Great Fish River by the 1770s. During the 18th century the Xhosa had been embroiled in two major civil wars over the chiefly succession. After both struggles, the unsuccessful contestants fled west across the Great Kei River, where they bore the brunt of the Xhosa wars against the Dutch and later the British. Various attempts to separate the colonists and the Xhosa were unavailing: in 1778, the Dutch decreed the Great Fish River as the boundary between the Xhosa and the Dutch, but Xhosa lived in the contested area to the west known as the Zuurveld, while trekboers were embedded in Xhosa territory to the east.

The establishment of the district of Graaff-Reinet in 1785 hardly improved matters. The area of magisterial jurisdiction was vast and its inhabitants unruly. Before the century was over, minor cattle raids had escalated into two frontier wars, the prelude to a struggle that lasted almost 100 years; the trekboers only expanded again after mov-

Principal peoples of southern Africa, 17th to mid-19th century.

ing north and outflanking the Xhosa. While the Dutch had superior firearms, the Xhosa had superior numbers, and both sides were internally divided. Thus, the first two frontier wars resulted in a stalemate, which ended only when the British acquired the colony permanently in the early 19th century.

By the end of the 18th century, then, when the British took over, the small Dutch East India Company outpost at the Cape had grown into a sprawling settlement in which some 22,000 whites dominated a labouring class of about 25,000 slaves and approximately as many Khoisan, as well as free blacks and "Prize Negroes"—slaves seized by the Royal Navy and reenslaved in the Cape—in Cape Town and a growing number of Xhosa in the eastern districts.

THE SLAVE AND IVORY TRADE

By the time the Cape changed hands during the Napoleonic Wars, humanitarians were vigorously campaigning against slavery, and in 1807 they succeeded in persuading Britain to abolish the trade; British antislavery ships soon patrolled the western coast of Africa. Ivory became the most important export from west-central Africa, satisfying the growing demand in Europe. The western port of Benguela was the main outlet, and the Ovimbundu and Chokwe, renowned hunters, were the major suppliers. They penetrated deep into south-central Africa, decimating the elephant populations with their firearms.

The more sparse, agricultural Ovambo peoples to the south also were drawn into the ivory trade. Initially trading in salt, copper, and iron and supplying hides and ivory to Portuguese traders, the Ovambo largely had been able to avoid the slave trade that ravaged their more populous neighbours. By the mid-19th century the advent of firearms led to a vast increase in the volume of the ivory trade, though the trade collapsed as the elephants were nearly exterminated by the 1880s. By then, traders from Angola, the Cape Colony, and Walvis Bay sought cattle as well as ivory. With the firearms acquired through the trade, Ovambo chiefs built up their power, raiding the pastoral Herero and Nama people in the vast, arid region to their south.

Neither Portugal's attempt to ban its nationals from slave trading in 1836 nor even the abolition of slavery in Brazil in the 1880s ended slavery in west-central Africa. Local merchants, chiefs, and elders turned to slaves to produce the tropical products demanded by Europeans and to serve as porters for the growing quantities of wax and ivory from the 1840s and '50s and rubber from the 1870s. Although the rubber trade was successful in the short term, excessive collection of wild rubber destroyed an irreplaceable natural resource, while new concentrations of population upset the ecological balance of a drought-prone environment.

Slave raiding in the east

British antislavery patrols drove the slave trade east, where ivory had been more significant. In the first decades of the 19th century, slave traders for the French sugar plantations in Réunion and Mauritius, who had previously drawn the majority of their slaves from Madagascar, turned their attentions to the Mozambique coast, while the demand from Cuba and Brazil also escalated. Thus, by the late 1820s Mozambique's slave exports were out-

stripping those of Angola, with demand from the French islands rivaling that of Brazil by the 1830s. The flow of slaves was augmented by turmoil in the interior of southern Africa and by slaves captured by the Chikunda soldiers of the Zambezi warlords; by the 1840s rival Zambezi armies were competing to control the trade routes to south-central Africa.

The most important area of slave raiding appears to have been in Malawi and northeastern Zambia. To the east of Lake Nyasa, the Yao—keen ivory traders from the 17th century—turned to slave raiding, obtaining firearms from the Arabs, subjugating the Chewa agriculturalists, and building up powerful polities under new commercial and military leaders. Displaced from northern Mozambique by the Ngoni in the 19th century, the Yao in turn pressured the Manganja peoples of the Shire Highlands. Although they never became large-scale slave traders, preferring instead to incorporate their captives, the Ngoni invaders added to the turmoil. While the first European observers probably exaggerated the extent of the depopulation, the political geography of the region was transformed as people moved into stockaded villages and towns and began to raid one another for captive women to work the fields while the men engaged in warfare. Vast numbers of people, especially women, were torn from their social settings, and earlier divisions based on kin came to matter less than new relationships between patron and client, protector and protected.

It was into this fluid and turbulent world that the first European missionaries to south-central Africa, inspired by the Scotsman David Livingstone, set up their Universities Mission in 1861. Although this mission ended in tragedy and failure, after Livingstone's death in 1873 other missionaries followed. In 1875 the Free Church of Scotland established the Livingstonia Mission in his memory, while the established Church of Scotland began work among the Yao at Blantyre the following year. From Lake Nyasa the Scottish missions spread inland to north-eastern Zambia and were followed by a large number of representatives of other Christian denominations in the last decades of the century. The missionaries and the Scottish-financed African Lakes Company that followed in their wake heralded the coming colonial challenge.

THE "TIME OF TURMOIL"

Given the turbulence caused by slave raiding in east- and west-central Africa, it is tempting to attribute to it also the unprecedented warfare in southern Africa in the second and third decades of the 19th century; the Mfecane, or Difaqane ("Crushing"), as this warfare is known, is currently much debated. As yet, however, there seems little evidence for extensive slave trading south of Quelimane until the 1820s, and the slave trade from Inhambane and Delagoa Bay remained paltry until 1823–44; the trade from these ports thus seems more a consequence than a cause of the wars.

Demand for cattle and ivory at Delagoa Bay seems rather more important in the emergence, by the late 18th century, of a number of larger states in the hinterland of Delagoa Bay. As in other Iron Age kingdoms, trade gave chiefs new ways of attracting followers, while elephant hunting and cattle raiding honed military organization. In the early 19th century, however, the number of European ships calling at Delagoa Bay appears to have contracted, and this may have increased competition for the cattle and ivory trade. Together with a series of devastating droughts, this competition may better account for the debilitating wars in which the larger northern Nguni chiefdoms in Zululand were embroiled by the second decade of the century. These battles occurred even before the rise of the Zulu king Shaka, whom an early historiography holds almost solely responsible for turmoil as far afield as the Cape Colony, Tanzania, and western Zambia.

Shaka

Shaka was thus the heir to, rather than the originator of, the intensified warfare in Zululand. Nevertheless, his military brilliance led to the emergence of the Zulu as the most important power in southeastern Africa. Within a few years Shaka had consolidated the numerous chiefdoms between the Tugela and Pongola rivers, although local and regional loyalties continued, and divisions within the royal family culminated in his assassination in 1828.

Initially, Shaka's most formidable rivals were the Ndwandwe, under the leadership of Zwide, who had driven Matiwane's Ngwane people onto the Highveld and Sobhuza's Ngwane north across the Pongola river, beyond the Zulu orbit. There, Sobhuza established the new conquest state of Swaziland (named for his successor, Mswati). In 1820 and again in 1823 Shaka defeated Zwide's armies, which broke into several groups. Zwide himself retired, but his generals fled northward. Clashing with one another and with the peoples in their path, the Ndwandwe (or Ngoni, as they became known) eventually established military states in northern Zimbabwe, Malawi, Zambia, and Tanzania, while the Ndwandwe general Soshangane established the extensive Gaza kingdom in south-central Mozambique. Adding greatly to the social dislocation of east-central Africa, Ngoni movements were dictated by the need to avoid more powerful African polities and to find new food resources after local cattle and crops had been exhausted through their raids. Within their military states, the Ngoni aristocracy monopolized cattle, incorporated the women and children of conquered peoples, and exacted tribute.

As in eastern Africa, where violence intersected with the intensifying activities of slave raiders, so in southern Africa the violence of this period needs to be disaggregated. Not only did warfare among the northern Ngoni precede the expansion of the Zulu kingdom; its rise does not suffice to explain the violence in the hinterland of the Cape Colony. There, the destructiveness of the settler presence was increasingly felt from the mid-18th century, as displaced groups of Khoisan and escaped slaves, carrying with them the commando system and the guns—and sometimes also the religion and the genes—of the white man, fled beyond the confines of the colony. In central and northwestern South Africa and southern Namibia these heterogenous groups of people, known variously as Basters, Griqua, Korana, Bergenaars, and Oorlams, competed for land and water with the Tswana and Nama communities and traded for or raided their ivory and cattle in the late 18th and early 19th centuries. By the 1800s the extension of the firearms frontier was disrupting the Orange River valley and intensifying conflict between the Sotho-Tswana chiefdoms beyond.

These disruptive processes were intensified when Britain occupied the Cape Colony at the turn of the 19th century. The displacement of Dutch East India Company rule by an imperial state in the early stages of its industrial revolution greatly expanded local opportunities for trade and increased demands for labour, just as the slave trade was abolished in the British empire.

This was of particular moment for the southern chiefdoms and rebellious tributaries attacked by Shaka as far as Pondoland. Many of the refugees fled either into the eastern Cape or west onto the Highveld, although their precise number is a matter of dispute. In both areas the arrival of the refugees added to upheavals of very different origin. The Mfengu, as the refugee population was known in the Cape, included in their ranks starving Xhosa victims of the 1834–35 frontier war, while the Mantatee or Fetcani (as the displaced population was known in the interior) were probably largely the product of labour raids by Griqua and Korana allies of frontier farmers.

Others shattered by the dual impact of the wars emanating from Zululand and the activities of labour raiders from the south scrambled to safety in the mountain fortresses of what is now Lesotho. There Mshweshwe, the Koena leader, built a new kingdom at Thaba Bosiu, defeating and then incorporating his main rivals through shrewd diplomatic marriages and his ability to provide them with protection. Mshweshwe quickly appreciated the utility of firearms and horses in the new warfare and of missionaries as diplomatic intermediaries.

Other dislodged Highveld peoples joined the Griqua polities along the Orange River or continued raiding along the Vaal and into the western Transvaal region, where the disorders prepared the way for the coming of Mzilikazi. Originally one of Shaka's commanders, Mzilikazi fled

Mzilikazi

from Zululand in 1823 with some 300 of his followers, known as the Ndebele (or Matabele). Over the next 15 years Mzilikazi created a 20,000-strong raiding kingdom in east-central South Africa by absorbing local Sotho-speaking peoples into his regiments. In 1837, harassed by his many enemies and defeated by expanding white farmers from the Cape Colony, Mzilikazi retreated across the Limpopo into southwestern Zimbabwe.

There, Mzilikazi established himself relatively easily, for the Shona polities were ill-prepared for the new form of warfare and were already weakened by the earlier incursions of the Ngoni and by drought. As in northeastern South Africa, the local populace was absorbed into Ndebele age-set regiments; a castelike society evolved, with the original Ngoni on top, Sotho in the middle, and Shona at the bottom. The relationships that the Ndebele established with groups beyond their immediate settlement ranged from friendly alliances to the regular exaction of tribute and random raiding. Beyond the range of Mzilikazi's armies, however, many Shona chiefdoms remained independent.

The Kololo

Yet another group dislodged by the warfare of this time, the composite Sotho group known as the Kololo, made its mark in west-central Africa. Defeated in warfare among the western Tswana, about 1840 Sebetwane led his followers across the Zambezi into northwestern Zambia. There they conquered the Lozi kingdom, which had been built up in the 18th century, and then dominated western Zambia. The Kololo triumph was short-lived, however; by 1864 the ravages of malaria, the accession of a weak and diseased king, and the revival of Lozi royal fortunes put an end to their hegemony. Nevertheless, a variant of Sotho is still the language of the region.

BRITISH OCCUPATION OF THE CAPE

During the Napoleonic Wars the Cape passed first to the British (1795–1803), then to the Batavian Republic (1803–06), and to the British again in 1806. The main impulse behind Britain's annexation was to protect its sea route to India. However, the British demands that the colony pay for its administration, produce raw materials for the metropole, and provide a market for Britain's manufactures and a home for its unemployed ineluctably drew Britain into defending the colonists, expanding their territory, and transforming the Cape's mercantile economy.

If the expansion of white settlement under the British led to a vast expropriation of African land and labour, it also led to a rapid expansion of unequal trading relations. Black-white exchange existed in the frontier zone from the early 18th century. British traders soon crossed colonial frontiers and were at Shaka's court by the early 1820s. They exchanged African cattle and crops for beads and brandy and on occasion may have purchased slaves, although even settlers well beyond colonial boundaries now disguised this as "apprenticeship" and "indenture."

Growth of missionary activity. From the end of the 18th century, European missionaries were crucial in the transformation of African society at the Cape. With Christianity came Victorian notions of "civilization" and "progress." "Progress" meant peasant production and entry into the labour market. The first converts in the Cape were the Khoisan in the east and north, and the Griqua, who by the 1820s had formed a series of independent if schismatic states in the Vaal-Orange confluence. The neighbouring Sotho-Tswana communities were also early sites of missionary activity. Two of the most famous 19th-century Scottish missionaries to southern Africa, Robert Moffat and Livingstone, worked among the Tswana. The most notable of the Tswana converts were the Ngwato, under the king Khama III (reigned 1875–1923), who established a virtual theocracy among his people, while in the eastern Cape the Mfengu were in the forefront of mission activity and peasant enterprise.

Impact of the missionaries

Initially Christianity tended to advance most rapidly among the disaffected and dispossessed, and especially among women; it was usually only after a major disaster undermined their belief systems that considerable numbers of men turned to the new religion. By inculcating individualism and encouraging the stratification that was to lead so many of their converts onto the colonial labour markets, the missionaries attacked much that was central to African society and developed an ideology to accompany colonial subordination. By the last quarter of the 19th century, missionaries and African evangelists of almost every denomination were working among the peoples of southern Africa, eroding chiefly authority and inculcating the new values and practices of the colonial world but also bringing new modes of resistance and educating many Christian Africans who later became outspoken critics of colonialism.

British development of the Cape Colony. In its constitutional development the Cape Colony followed the pattern set by Britain's other settler colonies in the 19th century. It was initially a crown colony governed by an autocratic governor, whose more extreme powers were modified by the presence in Cape Town of an articulate middle class and by the arrival in 1820 of some 5,000 British settlers. These groups demanded a free press, an independent legal system, the rooting out of corruption, and more representative institutions. After intense political struggle, Cape men were granted representative government in 1853, with a nonracial franchise that included a low property threshold, which, it was hoped, would defuse the discontent of both Afrikaners and the rebellious creolized Khoisan/Coloured population.

In 1872 the Cape gained full responsible government. The colour-blind franchise was retained but came under increasing attack. As a strategy for incorporating the more prosperous black peasants and artisans, it had been supported by white merchants, professionals, and officials. With the annexation of African territories and the creation of a mass black working class, however, it proved vulnerable, and in 1887 and 1892 the franchise qualifications were changed to restrict the number of black voters.

Initially imperial protection expanded Cape wheat and wine production, while the British did little to alter existing social and property relations. By the mid-1820s, however, imperial attempts to create a "free market" in labour—including the abolition of preferential tariffs and reform in the system of land tenure—had an explosive effect on the class relations of a colony dependent on slaves and serfs. New regulations ensured standards of treatment and established equality before the law for "masters" and "servants." Ordinance 50 of 1828, which ensured Khoisan mobility on the labour market, caused an uproar; in 1834 slaves were finally emancipated. Despite their formal equality before the law, newly emancipated slaves received only modest protection, from the handful of mission stations, against exploitative and often brutal conditions. By 1841, largely through "masters and servants" legislation, settlers had reimposed much of their old authority.

Despite the limited benefits for the underclasses, however, the British land and labour policies—together with a restructuring of local government—threatened many Afrikaners. Between 1834 and 1838, in a movement known as the Great Trek, parties of Voortrekkers ("Pioneers"), with their families and dependents, departed the Cape Colony. Their exodus was to become the central saga of 20th-century Afrikaner nationalism. Beyond the confines of the colony, they established separate republics in Natal, the Orange Free State, and the Transvaal, outflanking the Xhosa along the southeast coast, where the British were confronted by a series of interlocking crises.

The Great Trek

The first of these crises had erupted in 1799 shortly after the British first occupied the Cape. This was the third war between settlers and Xhosa in the Zuurveld and coincided with a mass uprising of Khoisan in Graaff-Reinet. Although peace was restored in 1803, the Xhosa remained in the Zuurveld until British troops drove them east of the Great Fish River in 1811–12; subsequent near-constant skirmishing again exploded into war in 1818–19, 1834–35, and 1846. For most of the century the Cape was dependent on British troops for its defense and for the further conquest of African territory.

By mid-century the western Xhosa were formidable foes, who used firearms and adopted guerrilla tactics. Thus, the eighth war (1850–53) was the most drawn-out and costly of all. As in 1799, a simultaneous uprising of Khoisan/

Coloured people at the Kat River settlement in the eastern Cape weakened the colonists' position. In the end, it was not British arms or settler prowess that defeated the Xhosa but internal tensions resulting from the activities of white traders, missionaries, and settlers. The Cape's northern frontier was now the Orange River, while in the east the land between the Great Fish and Great Kei rivers was appropriated for white settlement.

Xhosa millenarianism

In 1857 the internally divided Xhosa, exhausted by years of attrition, in the midst of severe drought and cattle disease, and undermined by the aggressive policies of the British governor Sir George Grey, turned to millenarian prophecies. They slaughtered their cattle and destroyed their crops in the belief that doing so would raise their ancestors from their graves and drive the whites into the sea. When the awaited salvation failed to materialize, some 30,000–40,000 Xhosa streamed across the frontier to seek work in the colony. An equal number died of starvation. Although Xhosa farther east fought the colonists again in 1877 and 1879, the slaughter of the cattle marked the end of Xhosa political and economic integrity. Thereafter, the annexation of the remaining African territories proceeded peacefully, if piecemeal. The last of the independent kingdoms to pass into Cape hands was Pondoland, in 1895.

THE EXPANSION OF WHITE SETTLEMENT

The Republic of Natalia and the British colony of Natal. The establishment of trekker republics in Natal and on the Highveld greatly expanded the frontiers of white settlement. The Voortrekkers, however, did not display any sense of national unity, and the parties soon fell out and set off in different directions. The trekkers enjoyed some spectacular successes as a result of their firearms, horses, and use of ox-wagons to form laagers (protected encampments), as well as strategic alliances with African chiefdoms; they found it far more difficult to establish permanent hegemony over the region.

Victory over the Zulu at the Battle of Blood River on Dec. 16, 1838, and divisions in the Zulu kingdom enabled the establishment of the short-lived Republic of Natalia, bounded to the north by the Zulu kingdom and to the south by the Mpondo. In 1843, however, the British, anxious to control the sea route to India, fearful of trekker negotiations with foreign powers, and concerned that trekker raids would spread to the eastern frontier, annexed Natal, leaving the Zulu kingdom north of the Tugela River independent until its disintegration in the civil wars that followed its defeat by the British army in 1879.

Through the 19th century, British Natal was surrounded by powerful African kingdoms and was heavily outnumbered by Africans within the colony. Constitutional development in Natal was slower and more erratic than in the Cape; colonists received responsible government only in 1893. Unlike the Cape, Natal never had a viable nonracial franchise: at the century's end few Africans had the vote despite the existence of considerable numbers of mission-educated black Christians. Racial practices in Natal—including the reservation of lands for African communal occupation, recognition of tribal authorities, codification of customary law, and control over urbanization through labour registration and influx control—were born out of the colony's weakness and provided precedents for 20th-century segregationist policies.

Racial practices in Natal

Absentee landowners bought up land claimed and vacated by the Voortrekkers and extracted rent from African producers, hoping increased white immigration would raise land prices. When in 1860 sugar was exploited successfully for the first time, indentured labour had to be brought from India to do the arduous work, because Africans—many of whom still had their own land and cattle—refused to work for the low wages offered on the plantations. By the last decades of the 19th century, however, land shortage and high taxes had forced large numbers of Africans to seek work on colonial labour markets.

Voortrekker republics in the interior. With the British annexation of Natal, most of the Voortrekkers rejoined their compatriots on the Highveld, where separate communities had been established in Transorangia and the western and northeastern Transvaal. Apart from a brief period in the mid-19th century, the British left them alone, controlling external trade and security threats through the coastal colonies. To ensconce themselves in the interior they fought major wars and established a series of accommodations with those Africans whom they were unable to conquer.

Compared to the British colonies, the racially exclusive republics between the Vaal, Hartz, and Limpopo rivers were weak outliers of the world economy, dependent on cattle ranching and hunting. Bitterly divided politically and ecclesiastically, these republics were unified in 1860 as the South African Republic, annexed as the British colony of the Transvaal between 1877 and 1881, and reconquered as the Transvaal during the South African War (1899–1902). The trekkers staked a claim to black lands, provided a framework for speculation and the beginnings of commerce, and established formal legal title to territory, though these claims were as yet barely effective. The incapacity of the settlers to wrest the indigenous inhabitants from their land resulted in the development of several types of labour coercion and control: slavery, clientship, indenture, debt bondage, and various forms of rent and labour tenancy.

The struggle to transform formal claims into actual land ownership and control continued well into the 20th century. Money was short, and government officials were paid in land, usually along with its African occupants. The settlers' accumulation of wealth was often the result of random looting and forcible, though sporadic, extraction of tribute, tempered by the limited physical capacity of the commando system. Surrounded by a horseshoe of powerful African chiefdoms, it was only in the last third of the 19th century, during a period of renewed imperial interest in the interior, that the balance of power shifted decisively in favour of white farmers.

The Orange Free State and Basutoland. Farther south, in Transorangia, a far greater proportion of the small settler community was tied to Cape and British markets through wool production. Of a population in 1875 of some 125,000, only the 26,000 whites had citizenship, but many European observers considered the Orange Free State, with its parliament and written constitution, a model republic. Despite the Dutch ancestry of the majority of the settlers, English was the language of commerce and education into the 20th century.

The existence of Mshweshwe's Basuto kingdom on the settlers' eastern flank meant constant friction. With the restoration of peace on the Highveld in the 1840s, many Africans attempted to return to their lands, only to find them occupied. Despite Mshweshwe's attempts to keep the peace, cattle raiding by his dispossessed subjects, together with increasing demands for land and labour from settler sheep-farmers, led to war in 1858 and again in 1865–69. On the first occasion, the Orange Free State was forced to sue for peace. On the second, Basutoland, internally divided and starved of arms by the British, was beaten. Some chiefs, especially in the north, offered their allegiance to the Afrikaners and, with their followers, became labour-tenants on their farms; others moved into the Transkei. In 1868, in response to repeated appeals from the Sotho, the governor of the Cape annexed Basutoland, leaving the Orange Free State in possession of the fertile Caledon River valley. In 1869 the frontiers of Basutoland were delimited, and shortly thereafter it was handed over to the Cape. In 1881, however, when the Cape government tried to disarm the Sotho, a war that the colony could not control broke out, and in 1884 Basutoland reverted to British rule.

War with Basutoland

The Orange Free State also constantly encroached on its western neighbours, the Griqua and southern Tswana states, which were also under frequent attack from the South African Republic. These attacks led to a growing alliance among the Tswana kingdoms and to protest from the missionaries and Cape traders, who feared the Afrikaners would block the main route to the interior. Nevertheless, the area came under colonial rule only after the discovery in 1867 of diamonds in Griqualand West.

Colonists in Angola and Mozambique. For much of the 19th century, Portuguese colonists in Angola and Mozam-

bique were fewer in number and weaker in authority than those in the interior of South Africa. At the beginning of the century, fewer than 1,000 settlers in each colony huddled on a number of estates around inland forts, along the Bengo and Dande rivers in Angola, and along the lower Zambezi in Mozambique. Most of them had intermarried with local peoples and were independent of Portugal. The metropolitan Portuguese were unable to control either the coastal trade or the activities of the merchants and warlords in the hinterland, who often acted in their name. Despite a mythology that held that the Portuguese, unlike the northern Europeans, did not differentiate according to race, from early times it is clear that whites had superior status and prestige—if not always greater power—in Angola and Mozambique. Although both territories gained somewhat from the Napoleonic Wars, it was not until the end of the 19th century that Portugal regained any of its colonizing energy.

Plantation agriculture

From the mid-19th century, Portuguese capital began to enter the colony and the Luanda hinterland, and planters experimented with raising coffee, cotton, cacao, and sugarcane, using the slaves who could no longer be exported. In the absence of an adequate administration or communications network, the plantations in Angola were never highly successful, although coffee cultivation spread among African peasant farmers in the region. The appropriation of African land for plantations was resisted, and Portuguese attempts to expand their colonial nucleus led to a series of wars with African peoples, followed by famine and epidemics. The instability of the last decades of the 19th century paved the way for the colonial period that followed.

Portuguese attempts to develop Mozambique met with even less success, given the lack of investment and prevailing disorder, as escaped slaves, soldiers, and porters formed bandit bands in broken country and attacked Portuguese settlements and African villages. In many areas domestic slavery underpinned the migration of young men to the labour markets of the south by the 1850s. Liberal governments in Portugal from mid-century were anxious to outlaw the feudal aspects of the *prazo* system but were unsuccessful, despite four military campaigns and a declaration in 1880 that the *prazos* were crown property.

In 1875 Portuguese rights to Delagoa Bay were recognized internationally. With the discovery of gold in the South African Republic the bay acquired a new importance as its closest outlet, and in 1888 Lourenço Marques became the capital of Mozambique.

MINERALS AND THE SCRAMBLE FOR SOUTHERN AFRICA

From the 1860s it was known that there was gold in the interior of southern Africa. In 1867 diamonds were discovered at Kimberley in Griqualand West to the north of the Cape Colony, followed shortly thereafter by discoveries of outcrop (surface) gold in the Transvaal and deep seams of gold on the Witwatersrand in 1886. The conjuncture of speculation in mining futures and land, the imposition of colonial or company rule, and an industrial revolution based on mineral extraction meant that the last third of the 19th century was one of the most traumatic in the history of the region. The language of racial domination, though hardly new, was now buttressed by social Darwinism and was particularly well-suited to an era of intensified land and labour exploitation.

Economic impact of the mineral discoveries. The mineral discoveries led to dramatic economic development. Roads, railways, and harbours were built. New coal mines were exploited. Manufacturing responded to the new markets, while the creation of an internal market for food was crucial in the commercialization of agriculture and the spread of African cash-crop production. Land prices soared, and the demand for labour became insatiable. Colonial conquest subjugated the remaining independent African societies and destroyed the bargaining power of black workers.

The diamond industry

Although most scholarly attention has focused on the gold mines, it was the diamond industry that pioneered many of the characteristics of southern Africa's labour control policies. People from all over the world came to Griqualand West to seek their fortune; between 1871 and 1875 more than 50,000 Africans from all over the subcontinent came each year, many of them lured by the prospect of purchasing firearms. Within a few years there was hardly an African chiefdom, from the Transkei to the Limpopo, that was not armed with guns, intensifying the instability of these years.

Initially, claims on the diamond fields were limited, technology was primitive, and small-scale black diggers could compete with whites. In the mid-1870s, however, chaotic production conditions, a flooded world diamond market, and labour shortages made the transition to larger units of production necessary. Joint stock companies were created, bringing international capital and a transformation of mining technology. By 1888 the thousands of claims of the previous decade had been monopolized by the De Beers Mining Company. For black and white workers the establishment of the De Beers monopoly was of immense significance. African migrant workers were now more rigorously controlled by pass laws, which limited their mobility, and by being confined to compounds for the duration of their work contracts. Many white miners lost their jobs or became overseers, and wages for all workers were sharply reduced.

The Witwatersrand

With the discovery of the Witwatersrand, attention switched from Kimberley to the South African Republic. The coastal colonies competed to control the lucrative Witwatersrand trade, and immigration mounted. When local capital proved inadequate, funds flowed in from Britain, Germany, and France. From the late 1880s gold outstripped diamonds as the region's most important export, and by 1898 the Witwatersrand produced about one-fifth of world gold output.

In 1889 the Chamber of Mines, an organization of mine owners, was formed to drive down the costs of production. This became even more important once deep-level mines were opened in the mid-1890s, because development costs were high, the ore low-grade, and the price of gold controlled. Skilled, unionized white workers from the mining frontiers of the world were able to protect their high wages, while the chamber formed two major recruiting organizations, the Witwatersrand Native Labour Association (Wenela) and the Native Recruiting Corporation, to extend, monopolize, and control the black labour supply throughout the subcontinent.

Throughout the region it was usually young men who were the first migrants, often sent by homestead heads, who tried to control their movement and their wages, or by chiefs who received a recruitment fee or a portion of the labourer's wage in tribute. For many young men, a period of labour migration could bring independent access to bridewealth. Although the process had its roots in the migration of Africans to colonial labour markets earlier in the century, migrant labour expanded after the mineral discoveries and had profound ramifications for the control of senior men over juniors and colonial administrators over taxpayers. Chiefs thus became increasingly anxious over their lack of control over young men and women and struck alliances with colonial administrators and recruiting agents to secure the return of migrants.

The scramble for southern Africa. The first move in the scramble for southern Africa came with renewed assertions of British supremacy in the interior. After much dispute, Britain annexed Griqualand West as a crown colony in 1871, transferring it to the Cape Colony in 1881. The multiple crises following the diamond discoveries led during the 1870s to failed imperial schemes to confederate the southern African territories, but imperial wars between 1878 and 1884 effectively ended the independence of the major African kingdoms. Of these conquests the best-known was the war in 1879 against the Zulu, which included a spectacular defeat of the British army at Isandhlwana; nevertheless, wars against the southern Tswana and Griqua, the Pedi of the eastern Transvaal, the western Xhosa, and the southern Sotho were the essential precondition for the creation of a unified South Africa.

The mineral discoveries whetted German imperial ambitions, and in 1884 Germany annexed the vast, sparsely populated territory of South West Africa (now Namibia).

The annexation challenged British hegemony in the region, raised fears of a German-Transvaal alliance, and accelerated the scramble for southern Africa. The possibilities of mineral wealth in the interior also revived Portugal's dream of uniting its African colonies. Portugal received short shrift from the other powers, however. At the Berlin West Africa Conference of 1884–85, Portugal secured the Cabinda exclave and a portion of the left bank of the Congo River on the Atlantic coast—considerably less than it claimed—and in 1886 the Kunene-Okavango region went to Germany. Portugal gained even less in Mozambique, which remained a narrow coastal corridor.

With the discovery of gold, the remaining independent African polities south of the Limpopo were conquered and annexed, and both within and beyond colonial frontiers concessionaires were spurred by prospects of further discoveries and the availability of speculative capital. The Limpopo constituted no barrier, and between 1889 and 1895 all the African territories south of the Congo territory were annexed. In south-central Africa the British competed with the South African Republic, Portugal, Germany, and Belgium, while in east-central Africa, to the west and south of Lake Nyasa, the thrust from the south encountered the less powerful but still significant antislavery missionary and trading frontier from the east.

Disasters of the 1890s

For many of the peoples of the subcontinent, the first phase of colonialism may have been overshadowed by the series of disasters that struck rural society in the mid-1890s, including locusts, drought, smallpox and other diseases, and a disastrous rinderpest epidemic that decimated African cattle holdings in 1896–97. Whereas before the colonial period such natural disasters would have killed large numbers in the short term but probably would have had little long-term consequence, the disasters of the 1890s drew considerable numbers of Africans into dependence on colonial labour markets for the first time and thus permanently changed the structure of African society.

The annexation of south-central Africa. From the 1860s it was known that there were "ancient gold workings" beyond the Limpopo, and by the mid-1880s Lobengula, the Ndebele king, was surrounded by concession-hunters. In 1887–88 the high commissioner at the Cape, fearful of Transvaal expansion northward, declared the region a British sphere of interest.

The story of how Cecil Rhodes came to South Africa to repair his frail health and stayed to become a millionaire on the diamond fields before he was 30 is legendary. In 1880 Rhodes entered the Cape parliament, and in the 1880s he played a key role in securing the British annexation of the Tswana kingdoms that straddled the road to the interior. One of the leading mine owners in Kimberley, by 1888 he had bought out his rivals and created the De Beers consortium. In 1890, when he became the Cape's prime minister, he was the most powerful man in southern Africa.

Rhodes hoped to find in south-central Africa a "second Rand" to outflank the South African Republic. In 1888 his agents secured exclusive mining rights from Lobengula for Rhodes' British South Africa Company (BSAC), which was granted a royal charter by the British government to exploit and extend administrative control over a vast area of southern and south-central Africa. Across the Zambezi, where the British were anxious to preempt European rivals, Rhodes engaged the newly appointed British consul for Malawi and Mozambique, Harry (later Sir Harry) Johnston, to establish his company's claims.

A flurry of treaty making in 1888–89 left BSAC with land and mineral concessions throughout present-day Malawi and Zambia. Despite the dubious legality of the treaties, the chiefs agreed to accept British jurisdiction over non-Africans in their domains and over external relations. In the European chancelleries, where the frontiers of Africa were being decided, the treaties played an important role in negotiations. In 1890–91 British, Portuguese, and German conventions established the frontiers of many of the modern states of southern Africa.

For Britain the BSAC's great advantage was its promise to make British occupation effective against contending European powers and bring capitalist development at minimum cost. In 1890 Rhodes sent a "Pioneer Column," consisting of 200 white settlers and 150 blacks, backed by 500 police, into Mashonaland; the real goal was the Ndebele kingdom, which was conquered in a deliberately provoked war in 1893. Although Matabeleland's conquest brought an anticipated boom in BSAC shares, by the end of 1894 it was clear that there was no "second Rand" in south-central Africa and that the future lay with the new deep-level mines coming into operation farther south.

As their hopes of discovering gold waned, settlers and the BSAC began expropriating African land, labour, and cattle. Settlers who participated in the war were granted lavish farms and mineral claims, both of which soon passed to speculative syndicates. A land commission perfunctorily set aside two reserves for the Ndebele on poor soils. In 1896 the Ndebele rose in revolt and were joined by a number of eastern Shona polities. Only the arrival of imperial troops and the collaboration of other Shona groups saved the company state.

These events left few resources for occupation north of the Zambezi until the late 1890s. Opposition from missionaries and the African Lakes Company ensured that the region around Lake Nyasa and the Shire River valley was separated from the BSAC sphere; it was declared the British Central African Protectorate in 1891, with Johnston as commissioner. Given the fragmentation and social divisions of the region, he found little difficulty in implementing a policy of divide and rule. Johnston's antislavery wars had the advantage of releasing labour for European employers. Wary of creating a landless proletariat, Johnston, like Rhodes, nevertheless believed that the protectorate's future development should be based on the marriage of white enterprise and black labour, assisted by Asian middlemen.

Lewanika

West of the protectorate, Africans were drawn more gradually under colonial rule, despite pleas from the Lozi king Lewanika that the British provide technical and financial assistance in exchange for mineral concessions, as promised in an 1890 treaty. Lewanika's scramble for protection in the 1890s was dictated by the same circumstances that initially had led him to invite whites into his kingdom in the mid-1880s. The 20 years following the restoration of the Lozi monarchy after the Kololo interregnum had been filled with civil war and succession disputes. By inviting the missionaries, and subsequently the BSAC, to Bulozi, Lewanika, like the Ngwato king Khama III to his south, hoped to bolster his internal position and gain the skills to enable him to deal with the intruders.

In 1897 the BSAC sent an administrator to Bulozi. Contrary to Lewanika's expectations, this spelled the end of Lozi independence. Despite Lewanika's "protected" status, over the next decade the powers of the king and the aristocracy were whittled away, and Lewanika's hopes to control the modernization of his state were not fulfilled. Bulozi became a protectorate within a protectorate, tied to the southern African political economy.

In northeastern Zambia, too, the process of imposing colonial rule came later, but in the end it was swifter and less violent than it had been to the south or east. The natural disasters of the 1890s diminished the ability of the more powerful groups to resist, while weaker peoples at first welcomed the end of Bemba, Ngoni, and Swahili exactions. A lack of resources spared the region major confrontations with colonialism: among Mpeseni's Ngoni, where gold was believed to exist, the onslaught was as dramatic as in Zimbabwe and the expropriation as brutal. Nevertheless, attempts to impose closer settlement, interfere with local agricultural techniques, and extract forced labour combined with natural disasters to produce extremely high morbidity and mortality rates in the early years of company rule.

Angola and Mozambique in the late 19th century. Although Portugal failed in its major territorial ambitions in the late 19th century, it nonetheless acquired about 800,000 square miles (2,000,000 square kilometres) of African territory, of which it controlled about one-tenth. In both Portuguese territories "pacification" became a sine qua non of economic development, and there were military campaigns or police actions in almost every year between

Resistance to Portuguese rule

Colonial southern Africa, 1884–1905.

From J. Fage, *An Atlas of African History*; Edward Arnold (Publishers) Ltd.

1875 and 1924, a measure of Portugal's weakness as a colonial power. The greatest resistance came from those people with the longest experience of Portuguese rule and with the necessary firearms.

The majority of Portuguese troops in both territories were black, a situation that turned every campaign into a potential civil war. Fragmentation of political authority, resistance of traditional elites threatened by colonial rule, and the precipitate introduction of taxes and forced labour policies also made resistance in the Portuguese colonies the most prolonged in early 20th-century Africa.

Colonial markets were of particular importance to Portugal, and tariff barriers were erected to protect its manufactures. Starved of capital and racked by financial crises, Portugal planned to develop the colonies by attracting immigration and foreign capital and by fostering plantation agriculture. In Mozambique, however, local employers could not compete with the Witwatersrand. By 1897 more than half the mineworkers on the Rand came from Mozambique, while thousands worked on South African farms.

Germans in South West Africa. The Germans were the last imperial power to arrive in Africa. Their annexation and control of South West Africa was eased by the intense cleavages that had opened up between the local Nama and the Herero chiefdoms, a result of their increasing involvement in the world economy during the 19th century.

Throughout the 19th century, displaced communities of Khoikhoi and Oorlams from the Cape had made their way into South West Africa. At first, they settled peacefully on land granted them by the local populace, some of them establishing mission communities. The advent in the 1830s of the Oorlam chief Jonker Afrikaner and his well-armed followers significantly altered the regional balance of power. Responding to an appeal from the Nama, who were being driven from their grazing lands by Herero expansion, Afrikaner settled at Windhoek. By gaining control over the all-important trade routes from Walvis Bay and the Cape Colony, he ensured Nama dominance over the Herero until his death in 1861. Wars between the Nama and Herero were exacerbated from the mid-19th century by the increasing cattle and ivory trade and the availability of firearms.

Initially Germany hoped to exploit the territory through a concession company, but it could not raise sufficient capital. The government was increasingly forced to intervene in local affairs, especially when settlers appropriated Herero cattle and grazing lands. The most formidable opponent of the Germans was Hendrik Witbooi, a Nama chief who tried unsuccessfully to unite the Herero and Nama against the Germans. After a lengthy guerrilla war, he was defeated in 1894.

The rinderpest epidemic, the alienation of the better-watered highlands, unfair trading practices, and increasing indebtedness led to an uprising by the Nama and Herero peoples in 1904–07. They were crushed in a genocidal campaign: the Herero population fell from about 70,000 to about 16,000, with many dying in the desert while

attempting to escape. The Nama were reduced by two-fifths. The handful of settlers had to turn for labour to the Cape Colony and Ovamboland, which was formally brought under colonial rule only when the South Africans took over South West Africa during World War I.

THE SOUTH AFRICAN WAR

If the Nama-Herero wars were among the most savage in colonial Africa, Britain fought an equally bitter, costly colonial war against the Afrikaner South African Republic. The reasons for the South African (or Anglo-Boer) War (1899–1902) remain controversial; some historians portray it in personal terms, the result of clashes between the president of the South African Republic, Paul Kruger, and the representatives of British imperialism, Rhodes and the high commissioner, Sir Alfred Milner; some argue that the British feared that the regional dominance of the South African Republic would open the way for German intervention in the subcontinent and endanger the sea route to India; others believe that the struggle was for supremacy over the richest gold mines in the world and the need to establish a state in the Transvaal that would fulfill the demands of the deep-level mine owners.

Even before the war, the South African Republic's inability to create and coerce a labour force was irksome to the deep-level mine owners, with their huge demand for labour and tight working costs. The liquor, railway, and dynamite policies of the South African Republic also angered the mine owners. Making use of a largely fomented clamour of British immigrants over their lack of voting rights, and secretly backed by the British colonial secretary Joseph Chamberlain, Rhodes plotted the armed overthrow of the republic by his lieutenant Leander Starr Jameson.

The Jameson Raid

The Jameson Raid in December 1895 was a complete fiasco, and the raiders were soon arrested. Rhodes was forced to resign from the premiership of the Cape Colony, and the alliance he had carefully constructed between English and Afrikaners in the Cape was destroyed. Previously loyal to the empire, Cape Afrikaners now backed Kruger against the British, as did their fellows in the Orange Free State. Nascent pan-South African Afrikaner nationalism was greatly spurred, and in 1899 a rearmed South African Republic issued an ultimatum to the British that amounted to a declaration of war. Over the next three and a half years, nearly 500,000 British troops were deployed against an Afrikaner force of 60,000 to 65,000. Some 6,000 British soldiers died in action and another 16,000 of infectious diseases. The Afrikaners lost some 14,000 in action and 26,000 in so-called concentration camps. The total number of African dead is unrecorded; according to low official estimates more than 13,000 died in the camps. In the end, the Afrikaners were forced to sue for peace, which was signed on May 31, 1902.

Even before the war ended, Milner had begun to "reconstruct" the vanquished Afrikaner republics; the most serious grievances of the mine magnates were removed, and an efficient bureaucracy was established. The smooth functioning of the mining industry was crucial both politically and economically. An acute shortage of unskilled African labour was resolved by the importation of 60,000 Chinese, despite the bitter opposition of white workers.

Africans were effectively disarmed and systematically taxed for the first time, and the pass laws were made more efficient. These changes also benefited white farmers, who were assisted in a variety of ways by the state. By 1906–07 the British were sufficiently confident of the new order they had established to grant self-governing institutions to male whites in the conquered territories, and in 1910 under the South Africa Act passed by the British parliament in 1909 the four South African colonies of Transvaal, Natal, Orange Free State, and the Cape were unified as provinces of the Union of South Africa.

The South Africa Act

SOUTHERN AFRICA, 1910–45

The nature of colonial rule. By the beginning of the 20th century the subcontinent was under European rule, and its disparate societies were increasingly meshed into a single political economy. The annexation of African territories meant the establishment of new states, and colonial rule was given perceptible effect by policemen and soldiers, administrators, tax collectors, traders, prospectors, and labour recruiters. Railroads connected the coast with the interior, opening up new markets and releasing new sources of labour. New boundaries were drawn that lasted beyond the colonial period, and the Zambezi became the frontier between the settler south and the "tropical dependencies" of East and Central Africa, although Nyasaland (Malawi) and Northern Rhodesia (Zambia) occupied a middle ground.

The exploitation of minerals, the capitalization of settler agriculture, and the establishment of manufacturing industries drew Africans into the world economy as workers and peasants, transforming class structures and political alignments and shifting the division of labour between men and women. Previously male occupations, such as hunting and warfare, declined. Indigenous production of nonagricultural commodities from cotton to iron suffered from the competition of cheap, mass-produced imports. The costs of colonialism were unequally distributed. In the areas of white colonization, the BSAC and the colonial powers supported the settlers. Elsewhere African ruling elites were able to strike compromises with their new overlords. On the reserves and protectorates of southern Africa, chiefs and hereditary headmen still controlled their following, although their authority was eroded as they became appointees of the colonial authorities.

Chiefs and headmen

Some blacks and whites were able to take advantage of economic opportunities developing in new towns and markets. Yet for the growing numbers of mission-educated Africans and Coloured and for Indian communities, the period was probably one of regression rather than advance. European racist ideology replaced an older tradition in the Cape of social dominance through economic control. Strident settler demands for urban segregation classified even wealthy Indian merchants as "uncivilized natives." Indian immigration into all the South African colonies was restricted, and in Natal a number of anti-Indian discriminatory measures followed the grant of responsible government in 1893. In the Cape, institutions became increasingly segregated.

Many Afrikaners also experienced a period of rapid change. In 1886 the South African Republic was still a preindustrial state controlled by a livestock-owning elite; by 1910 it was dominated by mining capital and formed the hub of the industrializing subcontinent. The injection of international capital, inflated land prices, the South African War, and imperial social engineering transformed Afrikaner society as painfully and perhaps more completely than African society.

Racial policies

Racially discriminatory policies were prompted by settlers' fears of competition from blacks and the growth of black class consciousness; they were given an intellectual underpinning by anthropologists and administrators fearful of rapid social change. The Portuguese espoused policies of African assimilation, yet obstacles to progress for the Afro-Portuguese and acculturated African elite were more rigidly enforced in the 20th century than they had been in the 19th. Thus, before 1945 the ideology of segregation was espoused by virtually all the governments of the region and by most whites regardless of political persuasion. For blacks segregation meant exclusion from citizenship; incorporation into a restricted and racially segmented labour market based on the use of migrant labour; government control of movement, urban residence, and trade union organization; the consolidation of the authority of the chiefs; and a recognition or invention of black ethnic identity in the African reserves.

The constitution of new colonial states and settler politics. In the new dispensation, whites, with state assistance, controlled private property and the means of production, while Africans were seen solely as labour. In South Africa after 1912 and the British colony of Southern Rhodesia (Zimbabwe) after 1923, settlers controlled the police and armed forces; elsewhere Africans manned the police and armies of the colonial state, although imperial troops remained the ultimate authority.

Divisions among whites

Settlers everywhere were united in their determination to assert white supremacy but were divided by class and

ethnicity. Particularly in South Africa, South West Africa, and Southern Rhodesia, political struggles among whites were often bitter. In South West Africa, German and Afrikaner settlers lived in uneasy tension, which increased in the 1930s when pro-Nazi demonstrations advocating a German takeover of the colony were common. In the Rhodesias, too, there was antagonism between British settlers and Afrikaners who made their way to the territory in the early years of the 20th century, as well as conflicts between the BSAC and white workers and farmers.

These political struggles were most intense in South Africa, which had the most developed economy, the largest and most diverse population (African, Indian, Coloured, and white), and the most acute class and ethnic differences. Class warfare between white workers and the mine magnates on the Rand was fierce until the 1920s. The years after the creation of the union were turbulent, with a civil war between Afrikaners when South Africa joined the British side in World War I and a series of mine strikes, which culminated in 1922 when members of the newly formed Communist Party of South Africa and recently proletarianized Afrikaners, who still dreamed of restoring their republic, joined ranks; they were defeated only after the declaration of martial law and a five-day battle between white workers and troops on the Witwatersrand that ended with 230 dead and the ringleaders hanged.

The South Africa Act of 1909 enfranchised white adult males, but, except for a diminishing proportion of black male property holders in the Cape, neither blacks nor women were enfranchised. Although white women received the vote in 1930, in 1936 Cape African men were removed from the common voters roll, and in 1956 the Coloured voters of the Cape were similarly disenfranchised. White men effectively were given control over the majority of blacks, and they retained this control until 1994.

South Africa was at the centre of Britain's southern Africa policies. Nevertheless, until the 1930s the union was poor, divided, and dominated by international capital. White settlers were Britain's closest allies. Although it overpowered its immediate neighbours, South Africa's expansionist ambitions in the region were largely blocked. Thus, in 1910 the union wished to incorporate Basutoland (now Lesotho), Bechuanaland (now Botswana), and Swaziland—three landlocked territories that had remained outside South African control. African and humanitarian opposition and Britain's desire for a foothold in the region prevented this incorporation, and the territories remained British protectorates. Until the mid-20th century, however, both Britain and South Africa assumed that the territories would ultimately become part of South Africa.

Although this did not happen, Basutoland, Bechuanaland, and Swaziland were locked into South Africa's economy. All three territories, which had been grain and cattle exporters, became increasingly dependent on the South African labour market, especially after South Africa implemented protectionist measures for white farmers. Administrators were often South African, and the form of indirect rule they practiced strengthened the authority of conservative chiefs, leaving little room for political progress.

South West Africa

The union was more successful in acquiring the vast colony of South West Africa, which it conquered from the Germans during World War I. Despite a League of Nations mandate that South West Africa be administered as a "sacred trust" for its indigenous inhabitants, South Africa's concern was to foster mining and to subsidize poor Afrikaner settlement in what was known as the "Police Zone." In 1917 Ovamboland was annexed; better-watered and therefore more densely populated, Ovamboland had long been able to resist dispossession. During the interwar years South Africa was able to defy the many resolutions passed by the League of Nations urging African social and educational advance, and the country continued to defy them even when the South African mandate was withdrawn by the United Nations in 1946.

South Africa also had designs on Southern Rhodesia. In 1922, however, when the British South Africa Company relinquished control of Southern Rhodesia, the predominantly British settlers opted for self-government under British rule, and the territory became a self-governing colony the following year. While British subjects of all races were enfranchised, high property qualifications excluded the vast majority of Africans, who formed 95 percent of the population, from voting. Between the 1920s and '50s, the governing party generally remained closely allied to the small group of mining companies that controlled the economy, while the opposition usually represented white farming and working-class interests.

In Nyasaland and Northern Rhodesia, self-government for the handful of whites was clearly impossible, although in both colonies settlers were given some representation. With the discovery of copper, the white population in Northern Rhodesia increased, but whites never achieved a political dominance comparable to that of their compatriots farther south.

Although copper mining was interrupted by the world depression of the 1930s, by the eve of World War II Northern Rhodesia was a major producer. In Nyasaland, the BSAC hoped settlers would develop the territory, but white immigration was restricted by Nyasaland's sluggish economic prospects. In both territories racially discriminatory policies protected the interests of white settlers. Nevertheless, the small numbers of whites and British proclamations of the paramountcy of African interests, differentiated these territories from those farther south.

Settlement in Mozambique and Angola

In Mozambique and Angola, too, settler numbers remained minute, despite Portugal's schemes to encourage colonial immigration. Before World War I, colonists consisted mainly of illiterate and unskilled peasants. Power remained in the hands of the governor-general, the highest colonial representative of the Portuguese government. In Angola the collapse of rubber prices in 1913 added to settler problems, and many went bankrupt; in northern Mozambique, campaigns against the Germans during World War I led to famine, forced labour, and high mortality from combat and disease. After the war, however, the colonies attracted new settlers as their economies recovered. In Angola, diamond production in the northwest was an additional stimulus for settlement.

The republican period in Portugal (1910–26) was accompanied by a flurry of activity among settler political groups, some of them in alliance with Afro-Portuguese and members of the creole elite angered by bureaucratic inefficiency and corruption. With the inauguration of Portugal's authoritarian "New State" in the early 1930s under Antônio Salazar, however, immigration schemes were dropped and strict vigilance was exercised over all political and economic activity in the colonies. Consultative institutions disappeared, and grand imperial rhetoric accompanied a return to protectionism, fostering Portugal's needs for cheap raw materials and a closed market.

Land, labour, and taxation. Everywhere in early 20th-century southern Africa the priority of administrations was for labour and revenue, and an extensive tax system was developed to address both needs: Africans' access to land determined the supply and cost of their labour, and, where land shortages did not suffice to push them into the labour market, taxation frequently did. In many areas the colonial state was weak, and colonial administrators feared rousing widespread resistance; efforts to collect taxes were often followed by flogging, hut-burning, and the confiscation of crops or cattle. Violencewas often most intense where administrations were weakest. In areas that had been under colonial rule for more extended periods, legislation forced Africans who had not already been dispossessed of their land into the labour market.

As long as Africans had access to land, however, they had some bargaining power. Money for taxes could be earned by increasing crop production or selling cattle. In many areas women did most of the farming, and young men worked periodically on white farms and in mines to earn money for cattle, fertilizer, seed, and plows. In the long run, however, Africans became locked into the money economy, and land shortage and indebtedness brought ever-increasing numbers into the labour market.

White agriculture and African reserves. At the beginning of the 20th century the vast majority of Africans in

the subcontinent still lived by farming, though in many areas they had become rent- or labour-paying tenants and sharecroppers on land claimed by settlers, syndicates, and speculators. During the 20th century these various forms of tenancy were transformed into wage labour, as white farms became increasingly capitalized.

Throughout the subcontinent, lands were reserved for sole African occupation by administrations fearful that total dispossession would lead to widespread African resistance. In South Africa, mining capitalists also came to see the utility of the reserves in subsidizing cheap labour: the limited agricultural production of the reserves supported the families of migrants, who could then be paid as single workers. On the other hand, white farmers wanted to take over the African reserves for their own use, eliminate competition from African producers, and reduce the employment status of Africans from tenancy to labour service. The Native Lands Act of 1913 and supplementary legislation in 1936 harmonized these conflicting interests, setting aside about one-eighth of South African land for 4,000,000 Africans, reserving the rest for about 1,250,000 whites.

The Native Lands Act of 1913

In Swaziland the 3,000 whites who had gained land as temporary concessions from the king in the late 19th century retained virtually two-thirds of the total in land settlements in 1908 and 1915. In response the Swazi royal family gained much popular support by establishing a national fund to repurchase the alienated lands, and by 1968 it had acquired almost half of the total. Basutoland, which had been deprived of its most fertile lands in the 19th century, was a de facto reserve although, as in Bechuanaland, land remaining in African hands was inalienable.

In Southern Rhodesia, too, where the BSAC developed commercial farming to attract immigrants and raise revenue, even the limited African reserves that had been set aside at imperial insistence were a subject of constant contention. The crucial legislation was the Land Apportionment Act of 1930, which barred African land ownership outside the reserves, except in a special freehold purchase area set aside for "progressive farmers." The best land was allocated to whites; less than one-third went to Africans. From 1937 Africans not required as labour on white-owned lands were removed to the reserves, which became increasingly congested.

Settler agriculture

Throughout the region white capitalist agriculture was possible only with extensive state support, which was not granted to Africans. The worldwide depression of the 1930s, which severely affected white farmers, intensified discrimination against African peasant production, and by the 1940s many rural areas were virtually dependent on migrant remittances.

Initially, similar policies were pursued north of the Zambezi. In Northern Rhodesia the colonial office attempted to increase settler numbers by opening the colony's best lands for white farming, while reserves were drawn up for African occupation. Although the reserves were large, like the reserves to the south, they were far from the railway line, contained poor soils, and were soon overcrowded and eroded. It was only after World War II that white land ownership was limited in Northern Rhodesia and attempts were made to address rural poverty.

In the Shire Highlands a handful of settlers owned nearly 34 million acres, while about one-eighth of all land belonged to the African Lakes Company until 1930, when it reverted to customary use. The plantations remained poor and inefficient until the 1920s and '30s, when tobacco and tea replaced coffee and cotton. Low pay, forced-labour practices, and squalid working conditions meant the plantations depended on labour tenants, sharecroppers, and migrants from Mozambique and the more marginal north.

African agriculture

South of the Shire Highlands, however, from about 1910 the administration began to encourage Africans to produce cash crops. Despite this evidence of African enterprise, however, racial bias and Nyasaland's poverty tended to handicap peasant agriculture. By the 1930s increased numbers of migrants sought work in South Africa and Rhodesia, especially after the 1913 ban on recruiting "tropical" Africans for the South African mines was lifted.

Despite South West Africa's mandatory status, Africans did not regain the land lost to the Germans; by 1946 the whites (who formed less than one-tenth of the population) controlled over three-fifths of the land in the Police Zone. Located on arid lands capable of supporting only sparse human and animal populations, the African reserves served as labour reservoirs, subject to police raids and pass laws. As settler farms on the better lands in the centre and south of the territory blocked older patterns of transhumance, independent pastoral production became difficult, and, while a few African families were able to retain their hold on rural resources, the majority were forced to seek work in town. Farther north, where the Ovambo retained control over their more fertile lands, restrictions on their access to markets meant that by the 1930s they, too, were increasingly seeking work in the colonial economy.

The invention of tribalism. In the areas reserved for sole African occupation, governments made use of African political structures, creating "tribes" where none had existed and governing through compliant indigenous chiefs and headmen. Imperial authorities at first sought to curb and undermine the powers of chiefs, whom they saw as the embodiment of their people and as potential leaders of resistance. Once the powers of the chiefs had been limited, however, fears of "detribalization" and the potential radicalization of African workers confronted administrations. In response, colonial governments throughout the region moved to bolster chiefs, granting them increased authority over their subjects while seeking to maintain their subordination to the colonial state and establishing local advisory councils as a substitute for popular enfranchisement and representation in central government. This creation of "tribal" institutions frequently created new identities and political interests.

Industrial development and increasing Westernization often made indirect rule through chiefs inappropriate to changing African needs, however. The extension of the market economy intensified divisions, especially as chiefs became identified with unpopular colonial policies and no longer had sufficient land for their commoner followers. The state recognition of chiefs, the imposition of "tribal boundaries," and land shortages meant that dissatisfied commoners could no longer check arbitrary rule by attaching themselves to alternative polities, as they had in precolonial times. Although urban migration provided some outlet, restrictions on African movement into the colonial towns, together with the often squalid living conditions and low wages, meant that moving to the towns was not an easy option.

Labour and the mining industry. At the beginning of the 20th century by far the strongest demand for labour came from the gold mines of South Africa. With the creation of the Union of South Africa there was for the first time a state strong enough to ensure the effective implementation of the laws and labour policies that had developed in Kimberley and on the Witwatersrand to control the workforce: migrant labour, compounds, revamped pass laws, and "masters and servants" legislation. The development of South Africa as the most powerful and industrialized country in modern Africa was built upon the labour of a poorly paid, mistreated, and disenfranchised workforce drawn from the entire subcontinent.

The early years of the century also saw intensified recruiting of African labour from Northern Rhodesia, Mozambique, and Nyasaland for the hundreds of small mines working scattered gold deposits in Southern Rhodesia. The difficulty of mining profitably in Southern Rhodesia meant that wages, food, housing, and health conditions were cut back ruthlessly, and mortality and morbidity rates were exceptionally high.

Across the Zambezi the absence of mineral wealth meant that Africans in Nyasaland and Northern Rhodesia migrated to the mines in Shaba (Katanga), Southern Rhodesia, and South Africa in search of money for food and taxation, although the opening up of the copper mines shifted some migrant routes to the Copperbelt. In the interwar years Northern Rhodesia and northern Nyasaland were no more than massive labour reservoirs.

In Angola and Mozambique, too, the economy was sustained by labour migration as the recruitment of labour

Migrant labour from Angola and Mozambique

for South African, Rhodesian, and German enterprises provided revenue for tax and trade. The Portuguese government attempted to control the flow of labour from Mozambique to the gold mines through a series of conventions with the Transvaal and later the Union government. Capitation fees were a major source of state income, while deferred pay ensured the migrant's return, tax payment, and purchase of Portuguese manufactures. Mozambique also benefited from a fixed proportion of the Transvaal's railway traffic. In a similar system in Angola, contract labour was sent to São Tomé; when this system was terminated as a result of allegations of slavery in 1908, the São Tomé planters also turned to Mozambique for labour.

Most Mozambican migrant labour came from the region south of the Save River; farther north, the Portuguese had granted wide mining, agricultural, and commercial concessions to chartered companies in the 1890s. Based on the old *prazos,* the chartered companies controlled more than half of the colony's lands. Under Salazar the concessions were allowed to expire, but this brought little respite. Southern Mozambique was entrenched as a labour reserve; elsewhere in the colony, as in Angola, Africans had to produce fixed quotas of cotton and rice.

The impact of migrant labour. It is difficult to determine the precise impact of migrant labour in southern Africa in the 20th century. In south-central Africa, for example, the major agricultural communities probably did not send migrants, and the majority of migrants usually came from areas already decimated by slaving and raiding. In other regions, earnings from migrant labour were often used, at least initially, to increase agricultural production, and many migrants maintained their links with the rural areas and retired there in old age. However, many Africans became dependent on the money economy and became locked into the migrant labour system; rural impoverishment resulted from the increasing congestion and soil erosion on the reserves. The division of labour in the countryside began to change, and the burden of agriculture fell increasingly on women and children.

The cheapening of black labour through migrancy rendered skilled white workers vulnerable to attempts by mineowners to reduce costs by substituting cheaper semiskilled black labour for expensive overseas workers.

The colour bar

Whites demanded a "colour bar" to protect their access to certain jobs. Initially formulated to reconcile white workers to Milner's decision to import Chinese labour, the colour bar was formally established in South Africa under the Mines and Works Act of 1911 and its amendment in 1926. At the same time industrial conciliation legislation introduced after the 1922 strike excluded blacks from the wage-bargaining machinery. These examples were followed in the Rhodesias as well, although on the Copperbelt white workers were weaker and the liberal impulse of government stronger, so that by the 1950s a skilled black workforce began to emerge there.

Urbanization and the development of secondary industry in South Africa and Rhodesia. Mining shaped southern Africa's experience of industrialization, but during the 19th century towns in the Cape and Natal engaged in small-scale manufacturing; this accelerated in response to the demands of the mining industry, although it was not until World War I that manufacturing made a significant contribution to the economy. By the end of the war the future of the mining industry seemed in doubt, while dispossessed rural Afrikaners began to enter the cities in search of work.

Although the plight of poor Afrikaners was frequently attributed to their refusal to do manual labour, they were at a double disadvantage in the towns. Unlike Africans—who had some access to the land—Afrikaners were totally dependent on their urban wage and lacked the skills of English-speaking workers. It was in response to this that the "civilized labour" policy, which favoured employers using white labour, was devised in the 1920s. The policy probably was more effective in spurring capital-intensive manufacturing and the employment of poorly paid Afrikaner women than in eliminating white poverty.

To meet the needs of Afrikaners in the cities, South Africa from 1924 promoted manufacturing through the levying of tariffs, the use of the gold tax to subsidize infrastructure development, the provision of inexpensive food to manufacturing workers, and the imposition of stringent controls to ensure low wages for black labourers. However, the insulation of South Africa's fledgling industries from international competition during the world depression and World War II may have been the most important factor in its economic expansion. Although Southern Rhodesia attempted some of the same strategies, its economy remained overshadowed by South Africa even after the establishment of the British Central African Federation in 1953. The development of manufacturing in South Africa and Southern Rhodesia led to a sharp increase in the number of urban Africans in both territories.

The African response. African peoples, who were so painfully drawn into the capitalist economy of southern Africa and were subjected to ever-increasing administrative, economic, and political control, did not acquiesce in their subordination without resistance. Most engaged in daily struggles to survive and devised individual strategies to resist exploitation. Yet they did not all experience their subjection in the same way, and to some extent this weakened resistance. The 20th century witnessed the rise of new classes, with the emergence of an African petite bourgeoisie and working class in the towns and a considerable degree of stratification in the countryside. Migrant labour both undermined and strengthened the authority of the chiefs, especially in areas where the colonial state was anxious to retain traditional structures for purposes of social control. Alongside the growth of nationalist movements among the educated elite and the organization of trade unions among workers, there was a continuation of royal family politics, a restructuring of ethnic identification, and a resort to millenarian solutions.

Royal family politics. In regions where large centralized states had existed at the time of the colonial takeover, royal politics continued to be of significance. In Barotseland, Swaziland, and Basutoland, where paramount chiefs were recognized by the British, the traditional aristocracy combined with the educated elite to protect their position and demand the redress of grievances. In both Matabeleland and Zululand, where the royal families had been militarily defeated, royalists combined to demand state recognition of the monarchy.

Political organizations and trade unions. Nonviolent African opposition to white rule—through the adoption of Western-style political organizations and the formation of trade unions—was longest and most intense south of the Limpopo, where the existence of substantial Coloured and Indian minorities gave an extra dimension to anticolonialism. In South Africa, between 1906 and 1913, Mahatma Gandhi formed the South African Indian Congress and led the first large-scale nonviolent resistance campaign against anti-Indian legislation. He gained limited success, although restrictions on Indian movement and immigration to South Africa remained in force. After his departure in 1914, however, the militancy of the Indian Congress was lost until after World War II. Nevertheless, Gandhi's example influenced later African nationalists.

The Coloureds of the Cape and Transvaal also mobilized politically in the first nationwide black political organization, the African Political Organization, founded in 1902, which sought to unite Africans in opposition to the South Africa Act of 1909. The formation of a separate Coloured Affairs Department to some extent diverted Coloured political energies from joint black action. Coloureds were prominent, however, in the All-African Convention, a body formed in 1935 that represented numerous African organizations. In 1943 the All-African Convention, along with several Coloured organizations, founded the Non-European Unity Movement, which rejected cooperation with the government and sought full democratic rights for all South Africans.

The All-African Convention

In 1912 educated Africans united various welfare associations, which had developed in the late 19th and early 20th centuries, into the South African Native National Congress (later the African National Congress [ANC]). They aimed to represent African grievances, overcome tribal divisions, and gain acceptance from whites through

self-help, education, and the accumulation of property. Demands for industrial education, individual land tenure, and representation in Parliament were accompanied by attacks on the pass laws, the colour bar, and the Native Lands Act of 1913; until the 1940s the ANC's methods remained strictly constitutional and appealed mainly to the educated elite.

The ANC had its counterparts farther to the north, partly because many early nationalists had either studied or worked in South Africa. Native associations and welfare associations evolved among the educated elite from the second decade of the 20th century and gave birth to congresses in Southern Rhodesia in 1934, Nyasaland in 1944, and Northern Rhodesia in 1948, all forerunners of more radical anticolonial movements.

Although Africans in South Africa were moving into industry by the end of World War I, their trade unions were hampered by pass laws, lack of recognition, and police harassment; strikes were illegal and often were put down with violence. Nevertheless, the period 1918–22 saw a great deal of working-class militancy, and in 1920 Clements Kadalie, a Nyasaland migrant, founded the Industrial and Commercial Workers' Union (ICU). Initially consisting of dockworkers in Cape Town, the ICU spread rapidly as a mass movement in the towns and in the countryside, where those who had been evicted responded with millenarian zeal to its message. At its height the ICU claimed 100,000 members and had branches as far afield as Southern Rhodesia and South West Africa, but by 1929 it had largely disintegrated. By that time the Communist Party of South Africa was organizing black workers. Black unions appeared elsewhere in the region after World War II.

From the early 1920s the South African government, seeking to preempt black radicalism, attempted to provide channels for the expression of African grievances through a variety of local consultative councils. In the Rhodesias and Nyasaland and, slightly later, in the smaller colonial territories, advisory councils, "tribal representatives" and rural "native authorities" played a similar role.

In Angola and Mozambique Africans had even fewer political rights, except for a brief republican period (1910–26) when political organizations, trade unions, and the press flourished. For a while it appeared that Africans and settlers in Angola would strive for similar reformist goals, but the Africans broke away to form organizations publicizing black grievances and demanding limited welfare and educational benefits. Crushed even before the advent of Salazar, these groups were revived as social and educational organizations, and it was only during the 1950s that they became overtly political.

Christianity and African popular religion. Even at their height political organizations and trade unions never reached more than a fraction of the African population, especially in rural areas. In many areas witchcraft eradication movements remained as sensitive a barometer of social distress: in 1933–34, for example, amid world depression, drought, and locusts, a cult offering adherents a medicine called *mchape* to deliver them from witchcraft swept central Nyasaland and eastern Northern Rhodesia. Antiwitchcraft cults and prophet movements drew on traditional religious and cultural beliefs, offering hope to a sorely pressed and poverty-stricken populace.

By the beginning of the 20th century, however, parts of South Africa had already experienced almost a century of Christian endeavour. The scope of mission work, already entrenched in the Shire Highlands and south of the Limpopo, was vastly extended as new societies appeared on the scene. The Roman Catholic church revived its presence in Angola and Mozambique and spread rapidly in the rest of the subcontinent. The consolidation of colonialism and the new challenges facing African society gave mission activity renewed vitality, and throughout the region black education and health remained largely the responsibility of Christian missions until after World War II.

Independent churches

At the same time, by the late 19th century many missionaries had come to oppose African religious leadership and practiced their own colour bar. Thus, many Africans turned instead to the independent churches that emerged in South Africa in the late 19th century and spread rapidly throughout the subcontinent.

African independent churches frequently stressed African political solidarity and religious autonomy and were often characterized by a millenarian vision that disturbed missionaries, settlers, and administrators. In some areas whites sought to suppress them. Although the break with a mission church betokened a desire for independence from whites, there were many motives for separatism. In Nyasaland and Northern Rhodesia, adherents of the millennial Watch Tower sect violently confronted state authority, while among the rural Shona and Kongo in Angola even the millenarian churches were more usually quietist.

The impact of World War II. Unlike World War I, World War II did not involve campaigns on southern African soil, although large numbers of black and white soldiers fought elsewhere in Africa. Yet in many ways this war had a greater impact. In South Africa manufacturing overtook mining and agriculture in its contribution to the economy, and large numbers of Africans settled in the major cities. In Southern Rhodesia, too, the war boosted the economy, and by its end tobacco farming and secondary industry had emerged as key economic sectors.

Economic expansion during the war led to increased organization among African workers, whose wages lagged far behind the rising cost of living. In South Africa these years saw a wave of African worker militancy, partly inspired by the Communist Party, and a reorganization of the African National Congress by a new younger urban constituency. The brutal suppression in 1946 of a strike by African mineworkers further radicalized many African nationalists and brought about a closer alliance between the ANC and the Communist Party.

African labour militancy

In south-central Africa, too, the end of the war brought an eruption of strikes, particularly a strike by railway workers in 1945, which led to the founding of a large number of African trade unions in Southern Rhodesia. In 1947 the British government dispatched a trade unionist to organize African mineworkers on the Copperbelt, while the first union in Nyasaland followed in 1949. With general strikes in Bulawayo and Salisbury in 1948, a new form of political action had emerged.

World War II was important in shaking up the politics of the region in other ways as well. Thousands of Africans had joined the army, and some came back home with widened horizons, while their experiences of demobilization and discriminatory compensation fueled nationalist feeling. The 1941 Atlantic Charter, which proclaimed the right of all peoples to self-determination, also stimulated political activists in southern Africa. In the 1940s the African National Congress began to demand full democratic rights in South Africa for the first time, and its influence, like that of the trade unions, began to be felt throughout the region, spread partly by returning migrant labourers.

For those territories under the authority of the British colonial office, the Colonial Development and Welfare Acts of 1940 and 1945 signaled Britain's commitment to the development of empire at a time of internal weakness. Thus, after the war Britain attempted to expand agricultural production through agricultural research stations, extension programs, promotion of technology, and conservation measures. These efforts largely benefited white estate owners rather than African peasants, however, and the attempted restructuring of peasant production prompted considerable rural unrest, providing anticolonial movements throughout the region with a large, disaffected constituency.

SOUTHERN AFRICA SINCE 1945

After the war the imperial powers were under strong international pressure to decolonize. In southern Africa, however, the transfer of power to an African majority was greatly complicated by the presence of entrenched white settlers. After an initial phase from 1945 to about 1958, in which white power seemed to be consolidated, decolonization proceeded in three stages: the relatively peaceful achievement by 1968 of independence by those territories under direct British rule; the far bloodier struggle for in-

dependence in the Portuguese colonies and in Southern Rhodesia (from 1965 Rhodesia, which achieved independence as Zimbabwe in 1980); and the denouement in South West Africa, which in 1990 achieved independence as Namibia, and South Africa, where the black majority took power after nonracial, democratic elections in 1994. While at the end of the colonial period imperial interests still controlled the economies of the region, by the end of the 20th century South Africa had become the dominant economic power. Despite the spread of multiparty democracy, violence, inequality, and poverty persisted throughout southern Africa.

The consolidation of white rule. Paradoxically, World War II and the rise of more radical African political movements initially consolidated white rule, as evidenced by the victory of the predominantly Afrikaner National Party in South Africa, the creation of the Central African Federation by Britain, and renewed white immigration to the Rhodesias, Angola, Mozambique, and South West Africa.

South Africa. Dissatisfaction with the wartime cabinet and fears of urban African militants lay behind the victory of the Reunited National Party (later the National Party [NP]), which ran on a platform of apartheid ("apartness") in the white elections of 1948. Although the NP won only a plurality of votes, its victory signified a new Afrikaner unity that resulted from 30 years of intense ideological labours and institution-building by ethnic nationalists intent on capturing the South African state.

The National Party

Although the various interests in the NP had different interpretations of apartheid, the party essentially had three connected goals: to entrench itself in power, to promote Afrikaner concerns, and to protect white supremacy. By 1970 these goals largely had been achieved. The NP controlled parliament, and many English speakers voted for the Nationalists—despite their declaration of a republic in 1960–61 and subsequent decision to remove South Africa from the British Commonwealth—believing that the NP alone ensured white domination. Economic and educational policies favoured Afrikaners, who became increasingly urbanized and less economically disadvantaged.

Under Hendrik Verwoerd, who served as minister of native affairs and later as prime minister (1958–66), apartheid took shape. Controls over African labour mobility were tightened, and the colour bar in employment was extended. From 1959 chiefly authorities in the rural reserves (renamed "Bantu homelands") were given increased powers and granted limited self-government, though they remained subject to white control. Ethnic and racial distinctions among whites, Africans, Coloureds, and Indians were more strictly defined and policed. Although Coloureds and Indians were subordinated to white rule and humiliated by racial discrimination, they nevertheless were privileged in relation to Africans.

Black opposition to apartheid policies in the 1950s was led by the ANC in alliance with other opposition organizations consisting of radical whites, Coloureds, and Indians. In 1955 this Congress Alliance drew up the Freedom Charter, a program of nonracial social democracy. Africanist suspicion of nonracialism and hostility to white Communists, however, led to the formation of the rival Pan-Africanist Congress (PAC) in 1959. Both organizations were banned after demonstrations against the pass laws in March 1960 at Sharpeville, in which police killed at least 67 and injured more than 180 African protestors, triggering massive protests. Increasingly draconian security legislation, the banning, exile, and imprisonment of leaders (including Nelson Mandela, the leader of the ANC), and the widespread use of informants resulted in a period of relative political calm in the 1960s.

The stability of the 1960s encouraged international investment, and the South African economy became far more centralized and capital-intensive. Economic growth made possible unprecedented social engineering, and the political geography of South Africa was transformed as millions of people were removed from so-called white areas to the black homelands. Access to welfare and political rights were made dependent on state-manipulated ethnic identities, which assumed new importance with the creation of the homelands. In 1976 the Transkei homeland was given "independence" by the South African government, and grants of "independence" followed over the next four years to Bophuthatswana, Ciskei, and Venda, though their "independence" was not internationally recognized.

South West Africa. In South West Africa, too, the National Party increased its control in the 1950s and '60s. Long governed as part of South Africa, in 1949 South West Africa became South Africa's fifth province, and its white population was swollen by about 3,000 immigrants. The economy grew dramatically, increasing the mobility of black workers and creating an urban-based black intelligentsia for the first time. Apartheid was extended to South West Africa, however, and in the mid-1960s its reserves were also consolidated into seven ethnically defined homelands under tribal authorities.

The small political associations in South West Africa after the war were profoundly influenced by their South African counterparts, but the first mass organization to protest against South Africa's policies was formed only in 1958; in 1960 this organization became the South West Africa People's Organization (SWAPO). Launched by Ovambo contract workers, SWAPO came to represent most black South West Africans in opposing apartheid, racial inequalities, and economic subordination to South Africa. After years of fruitless peaceful protest, SWAPO began a military campaign against the government in 1966.

SWAPO

Although South Africa did not recognize the authority of the UN, the issue of South African rule in South West Africa came before the UN regularly, and in 1966 the UN called for complete South African withdrawal. In 1973 the UN appointed its own commissioner for Namibia (as the territory became known in the 1970s); despite the presence of the UN commissioner and the intensification of SWAPO's military campaign, it was only after Angolan independence in 1975 and increasing international pressure that South Africa's policies began to change.

Angola and Mozambique. White power in Angola and Mozambique remained relatively weak in comparison to South Africa and South West Africa. After the war Portugal sought to maintain its colonies in the face of growing, if still slight, African urban nationalist movements by increasing the settler population dramatically. This was facilitated in Angola by a coffee boom and the discovery of minerals and petroleum and in Mozambique by government-instituted agricultural schemes.

These developments brought little benefit to the majority of Africans, however, who continued to work as ill-paid migrant labourers, their upward mobility blocked by settlers. Even in areas of limited fertility, Africans still had to produce their quota of cotton, rice, or coffee; most of the good land was taken over by wealthy white landowners and multinational companies, and the forced labour codes remained in operation until 1962.

Lesotho, Botswana, and Swaziland. The victory of the overtly republican National Party in South Africa challenged British interests in the subcontinent. The NP's economic policies appeared to threaten British investments in South Africa, while the Nationalists also renewed their demand for the incorporation into South Africa of Lesotho, Botswana, and Swaziland.

By the mid-1950s it was clear that the three High Commission territories could not be transferred to South Africa and had to be prepared for independence. Limited funds were made available for the provision of social services, education, soil conservation, and infrastructure development, but this assistance did little to reduce the territories' dependence on migrant labour to South Africa.

The Central African Federation. Alarm at the NP victory in South Africa also stimulated Britain into federating its south-central African territories as a bulwark against Afrikaner nationalism. Even before World War II, Northern Rhodesian whites had begun to consider federation with Southern Rhodesia as a response to growing African assertiveness, and support for federation increased after the war. At the same time, the growing importance of the copper industry in Northern Rhodesia attracted Southern Rhodesian whites to the idea of federation. It was widely assumed that Southern Rhodesia would provide managerial and administrative skills, Northern Rhodesia copper rev-

enues, and Nyasaland labour for the new entity. Africans in the north, however, feared that federation would prevent political advance and extend Southern Rhodesia's racist laws. Ignoring African opposition, in 1953 Britain's Conservative government brought the territories together in the Federation of Rhodesia and Nyasaland, commonly known as the Central African Federation.

Despite the rhetoric of multiracial partnership, the economic advantages of federation appeared mainly to benefit Southern Rhodesian whites.

Independence and beyond. The process of decolonization in south-central Africa and the High Commission territories was generally peaceful. By the late 1960s the few remaining nonindependent African countries were all in settler-dominated southern Africa. The 1970s were a time of escalating wars of liberation in Mozambique, Angola, Namibia, and Zimbabwe. The independence of the Lusophone colonies under self-proclaimed Marxist governments was crucial in shifting the balance of power against the remaining white minority states in the subcontinent. International involvement in the region increased, and by 1980 only South Africa and Namibia remained under minority rule.

For the territories of southern Africa, the continuance of apartheid in South Africa shaped the postindependence years; the liberation of these territories in turn inspired and politicized South Africa's black populace and transformed the balance of power in the region. In response, P.W. Botha, who became prime minister of South Africa in 1978 and led South Africa until 1989, massively increased defense expenditures and began a low-grade war on the neighbouring states, determined to destroy all ANC bases. South Africa destabilized the region by arming internal dissidents, who attacked schools, clinics, railways, and harbours. This intervention was especially devastating in Angola and Mozambique, but South Africa also destabilized eastern Zimbabwe and raided alleged ANC bases in Zambia, Botswana, Swaziland, and Lesotho.

Malaŵi and Zambia. By the late 1950s more militant national movements had emerged in the Central African Federation and were attempting to mobilize a disaffected peasantry in all three territories. The emergence of these nationalist movements profoundly disturbed the federal authorities. After sporadic unrest in Nyasaland in 1959 a state of emergency was declared, while in all three territories nationalist leaders were arrested and their organizations banned. The crackdown set off further disorder, and in the northern territories the British were persuaded to move toward decolonization. By 1961–62 the nationalists had been released and new constitutions drawn up, and in 1963 the federation was dissolved. In the following year the Malaŵi Congress Party under Hastings Banda and the United National Independence Party (UNIP) under Kenneth Kaunda won the first universal suffrage elections in Nyasaland and Northern Rhodesia, respectively, and led them into independence as Malaŵi and Zambia.

UNIP

In Malaŵi the unity of the nationalists was short-lived, and by the end of 1965 Banda had ruthlessly suppressed opposition activity and created a de facto one-party state. The press was muzzled, and Banda, together with an unpopular clique, controlled the party, parliament, and judiciary. In the early years of independence, Zambia also was plagued by internal divisions, although initially Kaunda handled these more circumspectly than his Malawian counterpart. Nevertheless, by 1972 Kaunda followed Banda's example and responded to political conflict and the electoral threat posed by his opponents by declaring Zambia a one-party state.

Economically, the differences between Malaŵi and Zambia were initially more striking than the similarities; Malaŵi was dependent on plantation agriculture, while Zambia depended on the export of copper. Both countries prospered in the first years after independence, but Zambia's economy declined dramatically with the collapse in the world price of copper. The weaknesses in Malaŵi's economy were only to become manifest a decade later.

Banda and Kaunda differed most, however, in their relations with the liberation struggles in the rest of southern Africa. In the hope of gaining control of northern Mozambique, Banda negotiated with the Portuguese and withheld assistance from Mozambican nationalists, who during the 1960s were beginning their military campaign. He also established close ties with the white South African government, which supplied much of Malaŵi's direct aid. Malaŵi thus became the foundation of South Africa's "outward-looking" foreign policy in Africa.

Although initially Zambia was as tied economically to Rhodesia and the Lusophone colonies, Kaunda backed the resistance movements there and supported UN sanctions against the white government in Rhodesia. He paid a heavy price. The sanctions closed Zambia's major trade and transportation routes through Rhodesia, and although alternate routes were established through Angola and new east-west lines through Tanzania were constructed by the mid-1970s, subsequent armed incursions from Rhodesia and South Africa and continued warfare in Angola and Mozambique disrupted the costly new trade and transportation lines.

By the end of the 1970s Zambia was one of the poorest countries in Africa. Poor peasants made their way to the towns, and this contributed to high levels of unemployment, social violence, and crime. Despite the end of the war in Rhodesia, which had been perceived as the main constraint on the economy, Zambia's economy deteriorated further in the 1980s. Elections were held in 1991, and the newly formed Movement for Multiparty Democracy under the leadership of the trade unionist Frederick Chiluba swept to victory.

During the late 1970s Malaŵi, long believed to have successful rural development policies, also faced economic crisis. The lean years of the 1980s saw a widening gap between rich and poor, which was worsened by Banda's support of the Mozambican insurgency movement Renamo and the influx of vast numbers of refugees from the civil war in Mozambique. By the 1990s an ailing Banda was confronted by a rising tide of popular and external pressure and was forced to allow multiparty elections, which were held in 1994 and won by the opposition United Democratic Front.

Lesotho, Botswana, and Swaziland. The independence of the majority of Britain's African territories put the independence of the High Commission territories in southern Africa on the British agenda, despite their continued economic dependence on South Africa and the relative weakness of their independence movements.

Lesotho, with high levels of literacy, was the first to organize. In 1952 Ntsu Mokhehle formed the Basutoland Congress Party, modeled on the ANC. In 1958 Chief Leabua Jonathan, who was to become Lesotho's first prime minister, founded the conservative Basutoland National Party (BNP), with the support of the South African government, the powerful Roman Catholic church, and the queen regent. Jonathan led the BNP to a narrow victory in the 1965 elections; Lesotho achieved independence in 1966. In Botswana and Swaziland, modern nationalist movements emerged somewhat later and were dominated by members of the royal families. In Botswana, which achieved its independence in 1966, Seretse Khama emerged as the first president. In Swaziland, where the presence of white settlers and South African and international economic interests held up full independence until 1968, the Swazi king Sobhuza II emerged as head of state through the overwhelming electoral majority of his Imbokodvo National Movement in the rural areas. Thus, in all three territories conservative governments anxious to avoid provoking South Africa emerged in the first elections after independence.

Jonathan and the BNP

Botswana was undoubtedly the most successful, economically and politically, and retained the most open political institutions and the most distance from South Africa. Dominated by a modernizing elite, the country's economy flourished with the expansion of cattle ranching and diamond, nickel, and copper mining. Botswana played a leading role in efforts to coordinate the regional economy. The BDP, with a primarily rural electoral base, ruled Botswana into the mid-1990s.

In Swaziland, Sobhuza II in 1973 declared a state of emergency and consolidated his rule after a more radical

opposition party showed strength in the 1972 elections. Until the death of Sobhuza II in 1982, all opposition to the government and its close links with South Africa was suppressed. In the 1980s and '90s political repression and competition for power within the ruling group intensified.

Fears that the more radical BCP would win the 1970 elections in Lesotho led Jonathan, supported by South Africa, to annul the election and suspend the constitution. Opposition leaders fled, and by the late 1970s chronic warfare had erupted in Lesotho's northeastern mountains. Through the 1960s and early '70s Jonathan was South Africa's most reliable regional ally, but he subsequently became an outspoken critic of South African policies. Jonathan's authoritarian rule continued until 1986, when he was deposed in a military coup supported by South Africa. Elections in 1993 were won by Mokhehle's BCP.

Angola and Mozambique. The longest, most divided, and bloodiest wars against colonialism in the subcontinent occurred in the Portuguese colonies. War first erupted in Angola in 1961. The initiative was captured by the urban-based Popular Liberation Movement of Angola (MPLA), under its poet-president Agostinho Neto. The MPLA was supported by communists in Portugal, the Soviet Union, and Cuba, but its hegemony was contested from the start by Holden Roberto's National Front for the Liberation of Angola (FNLA), based in Zaire, and by Jonas Savimbi's National Union for the Total Independence of Angola (UNITA), supported primarily by Ovimbundu in the south.

The MPLA and UNITA

In Mozambique the nationalist organizations were initially more successfully united. The anticolonial struggle was led by Eduardo Mondlane of the Mozambique Liberation Front (Frelimo), which was formed in 1962 by exiles in Tanzania. Internal dissent had been crushed by 1964, and Frelimo launched a guerrilla war against targets in northern Mozambique. Despite the assassination of Mondlane in 1969, a new phase of the war opened in 1971 under the leadership of Samora Machel, and by 1974 Frelimo controlled much of northern and central Mozambique.

Portugal's initial response to the outbreak of revolt in Angola and Mozambique was all-out war, and by the mid-1960s there were some 70,000 Portuguese troops in each territory. Large numbers of black troops were recruited, and villagers supporting the guerrillas were subjected to savage reprisals. In April 1974 the sheer cost of the wars—together with rising dissatisfaction with the government in Portugal—led to an army coup, the collapse of the Portuguese government, and Portuguese withdrawal from Africa.

When the Portuguese left Luanda in November 1975, Angola was in the throes of a civil war between its divided liberation movements. The war escalated as the United States aided the FNLA-UNITA alliance through Zaire and encouraged a South African invasion of Angola in 1974–75 in the hope of installing a pro-Western government. The Soviet Union supplied weapons to the MPLA, which was aided by Cuban troops. The South African invasion was repelled, but South Africa continued to destabilize the MPLA government over the next 15 years through its covert support for UNITA, which it hoped to install as its client. The MPLA eventually established control of Angola under Neto, but its government was undermined by South African incursions, the flight of most of the settlers at independence, incursions of Kongo peoples from Zaire (until 1978, when the FNLA dwindled in importance), hostility from the United States, and its own doctrinaire economic policies. By the early 1980s the MPLA had lost control over large areas in the south and southeast to UNITA.

Portuguese withdrawal also led to Mozambique's independence under a Frelimo government in June 1975, but the flight of skilled expatriates and Mozambique's proximity to hostile regimes in South Africa and Rhodesia caused immediate problems. The country was severely hit by a drastic cutback in recruitment by the South African Chamber of Mines in 1976, and, like Zambia, paid heavily for obeying UN sanctions against Rhodesia and for supporting the liberation movements. Nevertheless, in the early years of independence, Frelimo abolished many of the most hated aspects of colonial rule and greatly increased the availability of welfare resources for the black populace. Mozambican territory was raided by Rhodesia and South Africa in 1979, and this was followed by the infiltration of the Mozambican National Resistance (Renamo), a brutal insurgency group established by Rhodesian intelligence services in 1976–77.

By the late 1980s both Angola and Mozambique had jettisoned Marxist economic policies and had achieved more cordial relations with Western countries. In Angola peace accords signed in 1991 between UNITA and the MPLA roused widespread optimism, but war erupted once more when UNITA refused to accept its defeat in multiparty elections held in 1992. In Mozambique a peace accord was signed between Frelimo and Renamo in 1992; multiparty elections were held in 1994 and resulted in a victory for Frelimo. Resettlement of the more than two million Mozambicans uprooted by the conflict took place on a large scale during the mid-1990s, although the physical and psychological toll of the conflict remains incalculable.

Zimbabwe. African liberation in Rhodesia was closely tied to the independence struggles in Mozambique. The election of 1962—boycotted by African nationalists—was won by the extreme right-wing Rhodesian Front party (RF), which ran on a platform of immediate independence under white control. The Central African Federation was dissolved in 1963. Britain was unwilling to grant Rhodesia independence; in 1965 the RF, under the leadership of Ian Smith, unilaterally declared Rhodesia independent. Despite international pressure, Britain refused to use force against the illegal regime.

The banning of successive nationalist organizations and the detention and exile of their leadership led to fierce infighting and the emergence of two major liberation organizations, the Zimbabwe African National Union (ZANU), under Robert Mugabe, and the Zimbabwe African People's Union (ZAPU), under Joshua Nkomo. With Frelimo's military successes in northeastern Mozambique in 1971–72, and, more important, with the transformation of the power structure in the region after the independence of the Lusophone territories, a new guerrilla strategy began to make headway. Various attempts by the British to resolve the conflict—including a referendum on a new constitution in 1972—all failed, and by the late 1970s the Rhodesian army and the guerrillas pursued the war with increasing ferocity.

ZANU and ZAPU

By 1978 it had become clear that the Rhodesian government would not win the war, and Smith, under pressure from Western countries and South Africa, agreed in 1978 to allow the internal African opposition to contest multiracial elections the following year. These elections, however, excluded ZAPU and ZANU. Thus, despite the appointment of a black prime minister, the war continued unabated. In 1979 renewed negotiations in London ultimately led to a peace settlement that established majority rule, and in 1980 Mugabe and ZANU won a landslide electoral victory.

The release of a large number of unemployed, armed young men into the countryside bequeathed a violent legacy, and by 1982 the initial ZANU-ZAPU government coalition broke down in the face of increasing violence in Matabeleland, for which ZANU held ZAPU responsible. Early in 1983 Mugabe sent government forces to punish the people of Matabeleland. Despite the withdrawal of troops and an amnesty in 1988, memories of this brutal counterinsurgency campaign were even more traumatic than recollections of the liberation struggle.

Although the early years of Zimbabwean independence were economically promising, with the return of investment as sanctions were lifted and a series of good harvests, gross inequalities persisted. Despite its revolutionary rhetoric, ZANU (which ruled Zimbabwe into the mid-1990s) seemed more intent on replacing white government with black than with transforming the lives of the poor.

South Africa. For all the apparent success of its social engineering policies, by the late 1960s cracks had begun to appear in the National Party's edifice of control. It subsequently confronted multiple crises, as black opposition

The Black Consciousness movement

again broke to the surface with the emergence of the Black Consciousness movement in 1968, led by the charismatic activist Stephen Biko. The movement sought to raise black self-awareness and to unite black students, professionals, and intellectuals. As black political activity increased, the apparently monolithic NP began to fragment.

The economy also began to show signs of weakness by the mid-1970s. Inflation climbed steeply and the economy contracted; a reliance on imported technology contributed to a trade deficit. Whites, who constituted a declining proportion of the population, could not meet the demand for skilled and semiskilled labour. The small internal market and African trade sanctions also hampered growth.

Yet the economic growth of the 1960s had expanded the black working class and increased its confidence, and 1972–73 saw a wave of strikes and rapid growth of the trade-union movement. In some sectors the labour activism caused African wages to rise more quickly than white wages. Nevertheless, technological innovation led to high unemployment for the unskilled, and urban conditions for Africans continued to deteriorate as impoverished homeland inhabitants defied the pass laws and sought work in town. For them, the fiction of the independence of the homelands came to have a grim reality in the 1980s, as their homeland citizenship restricted their legal access to jobs and housing in the rest of South Africa.

The revival of labour activism and the independence of Mozambique and Angola further inspired the Black Consciousness movement. In June 1976 the government's determination to impose Afrikaans on black schools provided the flashpoint for prolonged countrywide protests, touched off after police fired on demonstrating students in Soweto (a black township outside Johannesburg). This event transformed political consciousness beyond the youth—although they remained in the forefront of protest thereafter—with far-reaching consequences. Churches were radicalized, large numbers of community organizations sprang up, and there was a resurgence of support for the banned ANC, particularly among young people.

In response, the government abandoned many aspects of orthodox apartheid: African trade unions were recognized, the pass laws were abolished, and attempts were made to co-opt the African middle and skilled working class (through the granting of limited urban and welfare rights) and to enhance the status of Indians and Coloureds (through constitutional change).

The reform process had stalled by the mid-1980s, and the state attempted to undermine black opposition by cultivating conservative African leaders, notably Chief Mangosuthu Buthelezi, head of the primarily Zulu Inkatha movement in Natal, which became the scene of internecine violence. When F.W. de Klerk ascended to the presidency in 1989, he faced continuing African militancy, international economic and cultural sanctions, renewed economic recession, and intensifying war in Angola and Namibia.

On Feb. 2, 1990, de Klerk announced his intention to free Nelson Mandela, lift the ban on many opposition parties (including the ANC and the PAC), and negotiate with the black majority for a new, nonracial constitution. Agreement on an interim constitution was reached in 1993, and in April 1994 Mandela was elected president of South Africa.

Nonracial constitution

Namibia. The independence of Angola prompted changes in South African strategy toward Namibia during the late 1970s, as South Africa attempted to transform the territory into a quasi-independent buffer against more radical change by proposing complex constitutional arrangements for a transitional government. The strategy, based on the co-option of a local black elite as a moderate alternative to SWAPO, was intended to placate international opinion while leaving control of Namibia in South African hands and keeping its military options open. The constitutional proposals were rejected by the international community, however, and in 1978 the UN Security Council passed Resolution 435, which set out proposals for a cease-fire and UN-supervised elections. South Africa did not move to implement this resolution, though it had accepted similar proposals earlier.

By the second half of the 1980s—in part because South Africa once more had been drawn into invading Angola—the war in Namibia was becoming increasingly costly for South Africa in military, political, economic, and diplomatic terms. Under joint pressure from the Soviet Union and the United States, South Africa finally agreed to implement Resolution 435, and democratic elections in 1989 were won by SWAPO, led by Sam Nujoma. In 1990 Namibia finally achieved independence.

Problems facing southern Africa. Despite the achievement of independence, the countries of southern Africa remained embedded in a regional economy consisting of enclaves of high capital investment and rural areas of increasing impoverishment. The new governments inherited authoritarian and often quite rudimentary states from the colonial regimes and had little experience of governing. Many countries faced the problem of creating a nation out of nationalism: the anticolonial movements had been led by a small, middle-class elite, mobilizing a largely peasant populace, and, with the transfer of power, older social, economic, and ethnic cleavages opened. These conflicts, though often described as resulting from age-old ethnic tensions, were as much the result of relatively recent historical patterns of regional differentiation. In many of the countries, particularly Angola and Zimbabwe, the legacy of the independence struggles was itself divisive.

Civil society—trade unions, civic associations, the judicial system—was frail throughout most of southern Africa, while in several countries the nationalization of major industries increased the power and patronage of a small bureaucratic elite. One-party states—whether nominally Marxist, as in the case of Angola, Mozambique, and Zimbabwe, or capitalist, as in the case of Malaŵi, Lesotho, and Swaziland—quickly emerged and paid scant regard to civil liberties. Disregard for the rights of the peasantry lay behind much of the tension in Mozambique and Angola, while in Zimbabwe land redistribution was hampered, and peasants and women who had been mobilized during the war became disillusioned.

In most countries the early years of independence were marked by advances in education and welfare, a major method by which postindependence governments secured popular support and distanced themselves from their colonial predecessors. Within a short time, however, the new governments began to falter. Declining world prices for raw materials undermined many southern African economies, while limited resources, a narrow tax base, and the cost of social programs contributed to high levels of indebtedness. The urban bias of policy makers, together with demographic growth and several years of drought, bedeviled most attempts at rural development.

Economic decline

After the first few years of independence much of the region experienced a decline in economic growth and a reduction in living standards. Even South Africa, the economic giant of the region, saw a dramatic drop in the standard of living and high levels of unemployment. Through the 1980s the International Monetary Fund and the World Bank imposed policies of structural adjustment with little positive effect, although the agencies forced the cutback of state welfare programs throughout the region.

Economic difficulties in the region were greatly increased by the continuance of liberation struggles, which imposed painful choices on the region: governments of countries that had won independence had to walk a tightrope between their natural sympathies for other nationalist movements and the danger that they would attract the ire of the more powerful settler societies. The majority of states did attempt, where possible, to support the liberation movements while protecting their own interests.

South Africa dominated the regional economy, accounting for about four-fifths of the value of the region's production in the early 1990s. The most important transport links and power lines ran from north to south, the major electronic media were based in South Africa, and the country continued to draw migrant labour from the region. Attempts by surrounding states to break South Africa's dominance through the formation of regional development organizations were only partly successful.

The end of the Cold War intensified the demand both internally and from the international community for polit-

ical change in southern Africa. The heightened interest in political reform was a major factor in the achievement of independence by Namibia, the end of apartheid in South Africa, the resolution of the conflict in Mozambique and attempts to end the civil war in Angola, and the conversion to multiparty rule in south-central Africa.

The 1990s thus seemed to hold the promise of a more hopeful future. These hopes, however, were extraordinarily fragile. Global economic restructuring, fluctuations in Western economies, and low commodity prices marginalized many of the countries of the region, while indebtedness and structural adjustment policies exacerbated internal tensions. Some two million people had been rendered homeless by the Angolan civil war, and the AIDS epidemic raged virtually unchecked, with infection rates soaring in some countries. The region was awash with weapons, haunted by the horrors of war, and bedeviled by intense poverty. Huge inequalities of wealth, dwindling resources, rapid population growth, high rates of illiteracy and disease, and political and ethnic divisions continued to plague southern Africa. The newly dominant ideology of market economics offered stark choices and little hope, while democracy remained extremely delicate. (Sh.M.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 945, 96/11, and 978, and the *Index*.

THE COUNTRIES OF SOUTHERN AFRICA

Angola

Angola is a country of southwestern Africa. Roughly square in shape, with a maximum width of about 800 miles (1,300 kilometres), Angola covers 481,354 square miles (1,246,700 square kilometres), including the Cabinda exclave, which is located along the Atlantic coast just north of the Angola-Congo (Kinshasa) border. Angola is bordered to the far northwest by Congo (Brazzaville), to the north and northeast by Congo (Kinshasa), to the southeast by Zambia, to the south by Namibia, and to the west by the Atlantic Ocean. There are 1,025 miles of coast, and the land frontiers total 3,023 miles. The capital is Luanda.

Angola is one of the largest oil producers in sub-Saharan Africa, and it is the largest of the Portuguese-speaking African states. Portuguese influences have been felt for some 500 years, but Angola acquired its present boundaries only in 1891. An anticolonial struggle from 1961 to independence in 1975 was followed by a continuing civil war that has left much of the country in ruins.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* From a narrow coastal plain, the land rises abruptly to the east in a series of escarpments to rugged highlands, which then slope down toward the centre of the continent. The coastal plain varies in width from about 125 miles in the area south of Luanda to about 15 miles in the Benguela region. The Bié (Angolan) Plateau to the east of Benguela forms a rough quadrilateral of land above the 5,000-foot (1,500-metre) mark, culminating at about 8,600 feet and covering about one-tenth of the country's surface. The Malanje highlands in the north-central part of the country are less extensive and lower in elevation, while the Huíla Plateau in the south is smaller still but rises steeply to an elevation of some 7,700 feet. The almost featureless plateau that covers the eastern two-thirds of Angola gradually falls away to between 1,650 and 3,300 feet at the eastern border. The highest point in the country is Mount Môco, near the city of Huambo, which reaches an elevation of 8,596 feet (2,620 metres).

Gerald Cubitt—Bruce Coleman Ltd.

Bedrock and laterite formations visible in the eroded landscape south of Luanda, in the subplateau region of Angola.

Drainage. The Lunda Divide forms a watershed on the plateau, separating north- and south-flowing rivers. In the northeast, rivers such as the Kwango (Cuango) flow out of Angola into the mighty Congo River, which forms the boundary between Angola and Congo (Kinshasa) for the final 90 miles of its course. The central part of the plateau is drained by the Kwanza (Cuanza), the largest river entirely within Angola's frontiers, which is about 620 miles in length. It runs for roughly half its length in a northerly direction before bending westward through a break in the escarpment between the Malanje highlands and Bié Plateau, and it flows into the sea about 40 miles south of Luanda. The southwestern part of the country is drained by the Kunene (Cunene) River, which heads south before turning west and breaking through the escarpment at the Ruacana Falls, after which it marks the boundary between Angola and Namibia to the Atlantic Ocean. Some rivers in the southeast of the plateau flow into the Zambezi River, which itself crosses the Kazombo salient in the far east of the country. Other rivers in this area feed the Okavango Swamp in Botswana. Small rivers in the south run into the internal drainage system of the Etosha Pan in Namibia, while others, often seasonal in nature, drain the steep western slopes of the escarpment.

Soils. The coastal plain consists of alluvia, chalk, and sand, underlain by oil-bearing formations occurring over the northern two-thirds. Crystalline bedrock of Precambrian age (*i.e.*, between about 570 million and 3.8 billion years old) emerges along the escarpment, and mineral deposits sometimes lie close to the surface. Much erosion has occurred in this area, and laterite formations are common. Most of the plateau in the eastern two-thirds of the country lies buried under deep deposits of infertile wind-blown Kalahari sands. River gravels of the northeast contain diamonds, and rare kimberlite pipes occur in this area.

Climate. Angola has a tropical climate with a marked dry season. The climate is largely affected by the seasonal movements of the rain-bearing intertropical convergence zone (ITCZ), the northward flow of the cold Benguela Current off the coast, and elevation. Rainfall is the key determinant of climatic differentiation, and it decreases rapidly from north to south and in proximity to the coast. The Maiombe forest in the Cabinda exclave receives the greatest amount of rainfall, about 70 inches (1,800 millimetres) per year, and Huambo, on the Bié Plateau, receives roughly 55 to 60 inches. In contrast, Luanda, on the dry coast, receives about 13 inches, while the far south of the coastal plain gets as little as 2 inches. The rainy sea-

Rainfall patterns

son lasts from September to May in the north and from December to March in the south. Droughts frequently afflict the country, especially in the south. Temperatures vary much less than rainfall. They generally decrease with distance from the Equator, proximity to the Benguela Current, and increasing altitude. The average annual temperature in Soyo, for example, at the mouth of the Congo, is 79° F (26° C), whereas in Huambo, on the Bié Plateau, it is 67° F (19° C).

Plant and animal life. When the Portuguese first arrived, parts of Angola were covered with dense rain forest, mainly in the north of the Cabinda exclave, the western edge of the Malanje highlands, the northwestern corner of the Bié Plateau, and along some rivers in the northeast. Much of this forest has been greatly diminished by agriculture and logging. Most of Angola's surface is covered with different kinds of savanna (grasslands with scattered trees), ranging from savanna-forest mosaic in the north to thorn scrub in parts of the south. Savanna vegetation is frequently accompanied by natural or man-made fires, and tree species are thus usually resistant to fire. True desert is confined to the Namib in the far southwest, which extends north from Namibia and is the home of a unique plant, the tumboa (*Weltwitschia mirabilis*), which has a deep taproot and two broad, flat leaves about 10 feet long that lie along the desert floor.

The fauna is typical of the savanna lands of Africa. The carnivores include leopards, lions, and hyenas, while the plant-eating animals are represented chiefly by elephants, hippopotamuses, giraffes, zebras, buffaloes, wildebeest, and different kinds of antelopes and monkeys. The land is rich in bird species and has a wide variety of reptiles, including crocodiles. The numerous insects include mosquitoes and tsetse flies. There are 13 national parks and nature reserves, but checks on hunting largely broke down with the spread of civil war. The giant sable antelope (*Hippotragus niger variani*) is found in the south and is particularly vulnerable. Other endangered species are the gorillas and chimpanzees of the Maiombe forest, the black rhinoceros, and the Angolan giraffe. Marine life is particularly rich in the south, because of the favourable effects of the cold Benguela Current, and it includes many species of temperate-water coastal fish.

Settlement patterns. The rural population is heavily concentrated in the highlands and along watercourses running off the highlands. The Bié Plateau alone contains about half the total rural population. In the north and centre of the country people live in villages, whereas in the south, where cattle keeping is important, there is a tradition of dispersed settlement and transhumance in search of pastures. A few !Kung live as nomads in remote areas of the far south. Decades of warfare have led to much enforced settlement in large villages.

Rural settlement

Urbanization stood at about 15 percent at the end of the colonial period, and it accelerated under the impact of civil war. More than one million people live in the capital city of Luanda, the major port and one of the oldest towns in sub-Saharan Africa. Farther south along the coastal plain, the historic town of Benguela and the port and industrial centre of Lobito are traditional rivals, while Namibe (Moçâmedes) is the port for the south and the country's largest fishing centre. Other important northern cities are Malanje at the eastern end of the Luanda Railway, and the coastal oil towns of Cabinda and Soyo (Santo António do Zaïre). Huambo (Nova Lisboa) is the regional capital of the Bié Plateau, and it is surrounded by a scattering of smaller towns, while Lubango (Sá da Bandeira) dominates the Huíla highlands.

The people. *Ethnic and linguistic composition.* Apart from a few Europeans and isolated bands of !Kung in the remote southeast, all Angolans belong to the Bantu linguistic family, which is found throughout central, eastern, and southern Africa. The largest ethnolinguistic group is the Ovimbundu, who speak Umbundu and who account for nearly two-fifths of the population. They inhabit the Bié Plateau, have migrated to Benguela and Lobito and areas along the Benguela Railway to the west and east, and live in fairly large numbers in Luanda. The next largest ethnic group is the Mbundu (or Akwambundu), who speak Kimbundu and who make up one-fourth of the population. They dominate the capital city and the Malanje highlands and are well represented in most coastal towns. Speaking Kikongo, the Bakongo in the far north are the third group. They account for about 15 percent of the population and are numerous in Luanda. Many Kikongo speakers are also found in neighbouring parts of Zaire and Congo. Lunda, Chokwe, and Ngangela peoples live scattered through the thinly populated eastern part of the country, spilling over into Zaire and Zambia. The Ovambo and Herero peoples in the southwest also live in Namibia, while the closely related Nyaneka-Nkhumbi people are confined to Angola.

Portuguese is the official language. It has become fairly widely spoken since independence, replacing Kimbundu as the lingua franca of Luanda, and it is understood in much of the country. English and Afrikaans are sometimes spoken in the south and east, especially by people who have migrated to Namibia and Zambia as workers or refugees, while French and Lingala are often understood among the Bakongo in the north.

Religion. The numerous and highly localized traditional animist religions of Angola are giving way to Christianity and syncretic Afro-Christian religions. The first centuries of Portuguese presence on the coast resulted in few conversions—apart from the early and atypical case of the Kongo kingdom—and it was only in the 20th century that Roman Catholic and Protestant missionaries began to have a major impact. The Roman Catholic church, to which more than two-thirds of Angola's Christians belong, is especially well established among the Ovimbundu. Various Protestant groups, with strong American and British links, make up the remaining Christians. Baptists have traditionally worked among the Bakongo people, Methodists among the Mbundu, and Congregationalists among the Ovimbundu. Afro-Christian religions, especially the Our Lord Jesus Christ Church in the World (Tocoist church), have spread from Zaire. Protestants and Tocoists were suspected of subversion in colonial times and suffered various forms of petty persecution. All religion was frowned upon by the Marxist-Leninist government in Luanda after 1975, and religious institutions lost their schools, clinics, newspapers, radio stations, and properties. However, freedom of conscience was legally guaranteed, and the Methodist church, which educated most of the government leaders, was relatively privileged. Other Christian denominations were closely watched and occasionally harassed, and the Jehovah's Witnesses were formally banned in 1978. Religious organizations regained considerable influence when the government abandoned communism, especially the Roman Catholic church, which played an important role behind the scenes in securing national reconciliation. Angola is unusual in Africa in having no Muslim population.

Status of religion after independence

Demographic trends. With only some 10 million inhabitants and a population density of about 21 people per square mile (8 people per square kilometre), the country is thinly populated, even by African standards. Vast areas in the semidesert coastal strip and the eastern two-thirds of the country are almost empty. War and famine are estimated to have killed about half a million people after 1975, but the population growth rate remains high. The birth rate is about average for southern Africa, but the high death rate has depressed the country's overall natural increase rate to a level well below average for the region. Life expectancy is low at 43 years for men and 46 years for women, and the population is predominantly young. Decades of warfare have resulted in about a tenth of the population living as internal refugees, congregating in the cities. It is estimated that about half a million people fled abroad during the anticolonial war, mainly Bakongo fleeing to Zaire and some Chokwe, Lunda, and Ngangela fleeing to Zambia. There was a renewed outflow of refugees in 1975, with the departure of more than 300,000 Portuguese and an unknown number of Africans. Many Bakongo refugees have returned to Angola, however, and there is even some immigration of Zairean Bakongo, who are attracted by employment opportunities in Angola's oil economy.

The economy. Angola had a diversified and rapidly growing economy in the late colonial period, but it suffered

badly after independence. A destructive civil war, the nationalization of most large enterprises, ineffective central planning, a heavily overvalued currency, and a constant exodus of skilled personnel were the main negative factors. From being the fourth largest coffee producer in the world, Angola sank to a position of complete insignificance. Manufacturing output slumped badly. Workers suffered a dramatic fall in living standards and were prevented from joining any kind of autonomous trade union. Angola was a net exporter of food before independence, but the country suffered from some of the worst famine conditions in Africa in the 1980s. Humanitarian aid has been hampered by poor security and political pressures. Economic reforms launched in 1988 remained largely ineffective, owing to the obstruction of the ruling party and the bureaucracy. The abandoning of Marxism-Leninism and the end of the civil war in 1991 offered new hope to a severely battered economy, which has the potential to be at the forefront of African development.

Growth of oil sector

The major exception to this dismal economic record has been the oil sector, which has made giant strides, boosting Angola to rank second as an oil producer in sub-Saharan Africa. Non-Portuguese foreign capital was exempted from outright nationalization, and, although the state took a share in the oil companies, management remained firmly in foreign hands. Moreover, the oil industry was protected from the worst effects of the fighting by its location on the coast and by the presence of Cuban troops. Oil accounts for more than 90 percent of Angola's foreign exchange earnings. Most of the rest comes from diamonds and fish, which have also remained mainly under foreign management. However, the linkage of the petroleum industry to the rest of the economy is minimal. The oil companies employ little local labour and do not reinvest their profits outside the oil sector. Nearly half of oil royalties have been spent on defense, and most of the rest has gone to pay for imports of essential foodstuffs for the urban population.

Resources. Angola's resources are considerable in comparison to those of most African countries, and they are particularly suited to the development of an industrial economy. There are large reserves of oil and natural gas, concentrated in the maritime zones off the Cabinda exclave and the Congo estuary. Total proven recoverable reserves of crude oil stood at more than 2 billion barrels in 1990. The quality of the oil is generally good, with a low sulfur content. Natural gas reserves are estimated at about 1.8 trillion cubic feet (50 billion cubic metres). Alluvial diamonds occur widely over the northeastern quarter of the country, with a high proportion of gemstones, and there are several kimberlite pipe formations that may be mined. Large iron ore reserves exist in the southwestern part of the country, but they are low-grade. Other minerals are known to exist in commercial quantities in Angola, especially in the area of the escarpment, but a great deal of systematic prospecting work needs to be done to gain a complete picture of the country's mineral resources. Angola's hydroelectric potential is one of the largest in Africa, estimated at more than 7,500 megawatts. Owing to the beneficial effects of the cold Benguela Current, Angola also has some of the richest fishing grounds in Africa, especially in the far south of the country. Stickleback, sardine, mackerel, catfish, mullet, and tuna are abundant, as are crabs, lobsters, and prawns. Timber resources are significant, with some 130 million acres of forest. The Maiombe forest in the north of the Cabinda exclave contains the most valuable commercial species, notably white tola (*Balsamiferum harms*) and limba (*Terminalia superba*). There are also stands of commercial timber along the rivers of the southeast, especially mussibi (*Guibourtia coleosperma*).

Agricultural resources

Fertile agricultural land is limited to a few favoured locations in the highlands and river valleys, and less than 10 percent of the land area is thought to be arable. The combination of poor soils and insufficient rainfall over most of Angola is a severe limitation to the growing of crops. However, the country's agricultural potential remains underutilized outside the Bié Plateau, the coastal oases, and the Ovambo floodplain on the Namibian border. Pastoralism is inhibited by tsetse fly infestation, poor pastures, and the lack of surface water in the Kalahari Sand zone. Conditions for pastoralism are best in the southwestern quarter of the country.

Agriculture and forestry. Prior to independence, a large estate sector dominated by Portuguese settlers coexisted with numerous small indigenous producers. Only about 3 percent of the land area was under cultivation. Of the cultivated area, less than 1 percent was irrigated. Coffee was king, and Angola supplied 19 percent of world coffee in 1974, with an annual output of more than 200,000 tons of green coffee in the early 1970s. Production was concentrated in the Malanje highlands and along the northwestern margins of the Bié Plateau and came mainly from the plantation sector. Cotton, consumed in local industries or exported to Portugal, was produced by both smallholders and planters in the Kwango valley and the coastal plain. The local market was entirely supplied by sugar intensively cultivated under irrigation by planters in the coastal oases. Oil palms, bananas, and other tropical fruits were associated with sugar on these plantations. Sisal was an estate crop cultivated mainly on the western slopes of the Bié Plateau, but its importance diminished in the 1960s, and it was often replaced by tobacco. Angola was a net exporter of food, with corn (maize) from highland African producers as the chief commercial food crop. Cassava (manioc) was widely grown as a subsistence crop by indigenous producers, together with millet, sorghum, beans, sweet potatoes, peanuts (groundnuts), rice, wheat, and potatoes. Cattle were concentrated in the southwest and were raised partly by traditional methods and partly on large modern ranches. Other livestock played little role in the commercial economy, but goats, pigs, and chickens were important for subsistence. Timber extraction from natural forests was concentrated in Maiombe in the Cabinda exclave and in Luso on the eastern stretch of the Benguela Railway. Some 123,500 acres of eucalyptus plantations along the western stretches of the Benguela Railway provided firewood for the steam locomotives and fed the paper-pulp plant near Benguela.

Agriculture after independence

This flourishing agricultural sector declined after independence. The nationalization of estates was followed by the creation of inefficient state farms. Ovimbundu contract labourers refused to work in areas where they were open to ethnically motivated attacks, notably in the coffee zones, and forced labour from the towns (the so-called voluntary brigades) proved a poor substitute. Smallholders were burdened with a centrally imposed system of cooperatives and by an inefficient public system of purchasing and distribution, which replaced the Portuguese petty traders who fled the country. The transport network deteriorated, insecurity spread throughout the country, the overvaluation of the currency acted as an increasingly heavy de facto tax on exports, and the collapse of manufacturing removed all incentives to sell to the towns. As a result, the urban population came to depend on food imported from abroad.

Agricultural collapse has been an uneven process. Subsistence crops have been little affected, commercialized food crops have done badly, and industrial raw materials and exported products have fared disastrously. Production of cassava and sweet potatoes actually rose slightly after independence, while the output of sorghum and beans fell by about 50 percent. Production of corn, bananas, and timber sank to about a quarter of their former levels, while only about a tenth of the sugar and beef output prior to 1975 was maintained. Coffee, cotton, and sisal plummeted to a mere 2 percent of former production levels.

Fishing. Before independence, there were about 700 vessels active in fishing, employing some 13,000 people, with an annual catch of about 300,000 tons. Namibe was the centre of this fishing industry, which stretched from Luanda in the north to the Bay of Tigres in the far south. The great majority of the catch was processed in modern factories and exported to Western markets frozen, canned, or as fish meal. A more traditional fish drying and curing industry supplied the local market and exported to a wider regional market before the anticolonial war began in 1961. Most of the fishing vessels, owned largely by the Portuguese, sailed away after 1975, packed with refugees, and the factories were destroyed or rusted away. Foreign

boats were then given licenses to fish in Angolan waters, on condition that a proportion of the catch be landed in Angolan ports. Some of the processing works were also renovated with the assistance of foreign aid. This arrangement kept the fishing industry alive at a time when much of the rest of the economy was collapsing. The catch fell to about three-quarters of preindependence levels, possibly because of declining fish stocks resulting from past overfishing or ecological changes.

Mining. Crude oil production is central to the economy, and production has nearly tripled since independence. Because Angola is not a member of the Organization of Petroleum Exporting Countries (OPEC), the country is not subject to any restrictive quotas on its exports. Angola enjoys promising geologic conditions, a high rate of exploration success, and relatively low operating costs.

Petroleum production

Production is concentrated off the coast of Cabinda, especially in the huge Takula field, and off the Congo estuary. Experiments show that deep-water operations in this zone may be profitable. There is some onshore production near Soyo and Luanda, and prospecting extends as far south as Kwanza Sul. Natural gas deposits have been found, but production is on a small scale. A state company was set up in 1977, under the name of the Sociedade Nacional de Combustíveis de Angola (Sonangol), to engage in joint ventures and production-sharing agreements, while leaving the management of the oil business largely in foreign hands. Chevron (United States) controls the Cabinda Gulf Oil Company, which is responsible for a little more than half of total oil output. Elf Aquitaine (France), Texaco (United States), and Petrofina SA (Belgium) are the other three companies producing oil in Angola, the latter exclusively from declining onshore oil fields. A host of other companies from around the world have been involved in prospecting for hydrocarbons, including companies from the United States, France, Italy, Spain, Portugal, Germany, Britain, Sweden, Norway, Brazil, and Japan.

Before independence, Angola was the fourth largest diamond exporter in the world in terms of value, with exports standing at 2.4 million carats, but output almost ceased after 1975, rose again in the late 1970s, and then crashed once more in the 1980s. The government nationalized the 77 percent stake held by Portuguese investors in the Angola Diamond Company (Companhia de Diamantes de Angola, or Diamang) and ran diamond mining in the northeast through the National Diamond Enterprise of Angola (Empresa Nacional de Diamantes de Angola, or Endiama), a parastatal company. Management problems were compounded by corruption, insecurity, and deteriorating relations with the Central Selling Organisation, controlled by the De Beers company of South Africa. In 1986, diamond mining was turned over to foreign enterprises under production-sharing agreements, and two years later the Angolans began negotiations with De Beers over diamond sales and technical assistance for the exploitation of kimberlite pipes. Output expanded again, and with the restoration of peace Angola is set once again to be a major player on the world diamond scene.

There are no other mining operations of any significance, although iron ore from the southwest was Angola's fourth most valuable preindependence export. But the Cassinga iron mines were heavily subsidized by the Portuguese, and their profitability was open to doubt. Production declined and ceased between 1975 and 1984, and the quality of the ore reserves is probably too poor to warrant the reactivation of the mines. Copper, manganese, gold, marble, black granite, and quartz have been mined or quarried on a small scale and there are plans to exploit phosphate deposits in Cabinda and Zaïre provinces.

Industry, construction, and utilities. Manufacturing, construction, and the hydroelectric industry were expanding rapidly prior to independence, but they were severely disrupted in the chaos that followed. Nationalization and the loss of skilled labour hit manufacturing especially hard, and UNITA guerrillas proved particularly effective at sabotaging electricity and water supplies. Huge debts accrued to the state, and factories operate on average at less than 30 percent of their capacity. Workers are paid partly in the goods produced by the factory (in spite of numerous official attempts to prevent this practice), absenteeism is rife, and productivity is abysmally low. Apart from some minor processing of primary exports, manufacturing is essentially import-substituting in nature and is concentrated in food, tobacco, cotton textiles, electrical goods, metals, vehicle assembly, oil refining, sawmills, cement, and construction for the oil industry. Most electricity comes from dams on the Kwanza, Kunene, Catumbela (Katumbela), and Dande rivers, at points where they breach the escarpment to reach the coastal plain. A large share of the country's installed generating capacity of more than 600 megawatts is out of use. Most of the nationalized manufacturing industries, construction businesses, and public utilities are being returned to the private sector, in some cases to their former owners.

Major manufactured goods

Finance. Banks were nationalized after independence. The National Bank of Angola (Banco Nacional de Angola) acts as central bank, bank of issue, and commercial bank, while the People's Bank of Angola (Banco Popular de Angola) acts as deposit bank. Foreign banks began to reenter the country in 1985, but banking remains overwhelmingly in the hands of the state. Most savings are held in informal banking structures outside the cumbersome state system. Foreign investment is highly concentrated in oil, diamonds, and fishing, but it is set to spread through the economy as liberalization proceeds and nationalized assets are returned to the private sector.

Trade. Oil accounts for more than 90 percent of exports; about two-thirds goes to the United States, where the low sulfur content of the crude oil is appreciated by refineries. The economy is thus highly vulnerable to shifts in the price of oil. After independence there was an attempt to obtain a growing proportion of imports from the communist bloc, and even to join Comecon. Huge quantities of arms were obtained, but otherwise this strategy met with little success. Civilian imports come mainly from Portugal, France, Brazil, and the United States. In addition to machinery, vehicles, and pharmaceuticals, Angola imports large amounts of food and raw materials, nearly all of which could easily and cheaply be produced locally. Even raw cotton is imported. While Angola's trade in commodities has remained in the black, the balance of payments has generally been in deficit, leading to the accumulation of arrears and the growth of the foreign debt. The return of peace should bring down imports of weapons and food, while the devaluation of the highly overvalued currency should boost exports.

Transportation. Angola achieved independence with an excellent transport network for an African country so large and thinly populated, in part because of Portuguese military imperatives after 1961. This sector of the economy has suffered more than any other from the effects of war and the lack of maintenance, and a rehabilitation plan, backed by foreign aid, was launched in 1988.

A grid of tarred roads links the major geographic centres of economic activity; in the late 1980s there were about 23,000 miles of paved road. They run into Zaire and Namibia and come close to the Zambian frontier. The complete road network totaled more than 46,000 miles in 1973. During the civil war, travel on the roads was almost impossible except in convoy with an armed escort, and numerous bridges were destroyed. Bus service between towns has ceased, and public transport within towns has almost ground to a halt. Taxis are rare.

Railways

The Benguela Railway is the longest of the country's railways, extending from Lobito on the coast to the Zairean frontier, and is owned in part by Tanks Consolidated Investments, a subsidiary of the Société Générale de Belgique. It provides the shortest route to the sea for the Zairean copper mines, on which the railway's profitability depends, but it has not functioned to the east of Huambo since the civil war began and has often been completely out of use. The Luanda Railway, which was nationalized in 1918, depends on coffee and cotton for its traffic, while the Namibe Railway, which was owned by the state from the outset, depends on iron ore. Both railways have functioned only episodically since independence. The short private Amboim Railway in Kwanza Sul has been out of action since the collapse of the plantations it served.

Lobito is the finest and best-equipped port in the country, but it has been underutilized since corn exports from the Bié Plateau and mineral traffic from Zaire ceased, with traffic reduced to about one-fifth of preindependence levels. Luanda has a good harbour, but it has been poorly managed, long delays are experienced, and it handles less than half the cargo it did before 1975. Since iron ore shipments have stopped, Namibe's activity has been based essentially on its role as the country's major fishing port. In contrast, Malongo and Soyo have grown in importance with the oil boom, although they are much poorer natural harbours. The lower Congo River is used by seagoing vessels up to Nóqui on the Zairean frontier, and small craft ply the lower Kwanza for about 140 miles. The merchant navy consists of several cargo vessels, unadapted to container traffic, which are run by a parastatal company. Most cargo is carried on foreign ships.

Travel by air became the only safe means of transport during the civil war, and the network of airfields left by the Portuguese has been intensively used. The number of passengers carried has more than quadrupled since 1975. The national airline, TAAG (Linhas Aéreas de Angola), has a fleet of 20 aircraft. Foreign airlines land only in Luanda.

Administration and social conditions. *Government.* Colonial rule was exercised by a right-wing dictatorship in Portugal from 1926, and the transfer of power at independence took place without any form of election. The Popular Liberation Movement of Angola (Movimento Popular de Libertação de Angola; MPLA) seized power in the main cities by force of arms, and the constitution of Nov. 11, 1975, as amended in October 1976, established a one-party state modeled on those of eastern Europe. The MPLA officially became a Marxist-Leninist vanguard party in 1977 and moved into close dependence on the Soviet Union and Cuba. The latter provided troops to combat a guerrilla challenge mounted by the National Union for the Total Independence of Angola (União Nacional para a Independência Total de Angola; UNITA). Faced with an endless war and the collapse of most of the economy, in 1985 the MPLA initiated a timid process of economic reform and began to turn more to the West. Cuban troops began to withdraw in 1989, following South African acceptance of Namibian independence in December 1988, and in 1990 the MPLA abandoned Marxism-Leninism. The civil war ended in 1991 with an agreement to introduce a new constitution. The political system is now based on the principles of the guarantee of full human rights and multiparty elections for the presidency and for parliament. However, in order to prevent the emergence of ethnically based parties, all political parties must prove that they have support in a majority of the country's 18 provinces before they can compete in elections.

Political movements

Education. After decades of neglect, the Portuguese began a crash program of education in 1961, resulting in a literacy rate of between 10 and 15 percent at independence. The MPLA aimed to achieve primary education for all after independence and tripled primary school enrollment between 1976 and 1979. However, this was followed by a halving of primary school enrollment during the 1980s. Conditions in schools declined dramatically, with an acute shortage of teachers and a lack of the most basic teaching materials, including books. Enrollment in secondary schools and in the university, which was founded in 1962, expanded continuously after 1975, as these institutions suffered less than primary schools from political insecurity. But there was a severe lack of teachers and teaching materials, and most faculties in the university have been closed for long periods because of alleged political agitation. It is estimated that recruitment into the armed forces of the MPLA and UNITA has had a greater impact than the formal educational sector on the spread of literacy, the increased use of Portuguese, and the acquisition of technical skills. Many Angolans have also been trained abroad, especially in Cuba and the Soviet Union.

Health and welfare. As with education, the Portuguese made a major effort to "win hearts and minds" after 1961 by expanding health and welfare programs. The MPLA government came to power with even more ambitious schemes, but initial successes were followed by an almost complete collapse of services, especially in the rural areas. Doctors and other medical personnel have been particularly prone to flee abroad, and they are reluctant to work in remote and dangerous parts of the country. Medicines are in short supply. Malaria and severe malnutrition are rife, and cholera epidemics are a frequent occurrence. Infant mortality rates are among the highest in the world. Urban housing schemes have been submerged by the flood of refugees from the countryside. Social conditions and the health situation in Luanda's slums have been worsened by acute shortages of water. Unemployment, inflation, empty shops, and the collapse of public transport have hit the slum dwellers hard, while the political and bureaucratic elite have benefited from a network of special shops, good housing, and other advantages financed from the proceeds of the oil economy.

Social conditions

Cultural life. Angolan culture is fragmented into a myriad of highly localized clusters, but its origins lie within a broad central African Bantu tradition, shared with neighbouring countries. Angolan peoples generally trace their descent through their mothers, and lineages or extended families are important in social life. People often give some allegiance to precolonial kings and chiefs, even though their political functions have long been in abeyance and they may be situated on the other side of a modern political frontier. There are also strong vestiges of a widespread precolonial system of slavery, which the colonial authorities were slow to abolish. Polygamy, the veneration of ancestors, and combating witchcraft are practiced, even by many converts to Christianity. Wood, clay, copper, reeds, ivory, shells, and the human body are the main media for the decorative arts. The wooden sculptures of the Chokwe people, the carved ivories of Cabinda, and the elaborate hair styles of the Nyaneka and Nkhumbi peoples are especially famous. Music and dancing play a central role in cultural life, with drums as the basic instrument, and there is a rich oral literature. There have been attempts to preserve this cultural heritage, which remains largely outside the domain of formal cultural institutions, but political instability and corruption after 1975 led to tragic losses of ethnographic collections.

Traditional culture has been increasingly overlaid by Western influences, which tend to predominate in the towns. The 19th century saw the emergence of a dynamic group of educated Creoles in Angolan towns, similar in some respects to those of Sierra Leone, who wrote newspaper articles, history books, novels, and poems in Portuguese. The right-wing dictatorship in Portugal drove much of this literary activity underground after 1926 but failed to destroy it altogether. The leader of the MPLA at independence, Agostinho Neto (1922–79), was famous for his poetry throughout the Portuguese-speaking world. The MPLA government exercised a rigorous system of censorship, and the output of the officially sponsored Union of Angolan Authors since 1975 has been disappointing. This was to some extent offset by the emergence of a national television service and the beginnings of a national film industry. An ambitious program to expand museums, libraries, and archives bore little fruit. Indeed, many fine collections built up in colonial times were destroyed, dispersed, or made unavailable to the public.

Effects of Western influences

Sports are completely dominated by football (soccer), which is a national passion and is played by people of every social stratum. Some Angolans have become footballers of distinction, but they tend to play in the clubs of Portugal and other European countries, where conditions are more attractive.

For statistical data on the land and people of Angola, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The Stone Age population of Angola was very sparse. Ironworking and settled agriculture spread during the 1st millennium AD, but hunting and gathering remained important for centuries. Bantu languages may have entered Angola with iron technology, and they were certainly spoken by the 15th century.

The slave trade. Portuguese navigators reached the

Kongo kingdom in the northwest in the 1480s and converted the king of Kongo, together with many Bakongo, to Christianity. Rumours of great silver mines in the interior and the growth of the slave trade led to a more aggressive policy. Paulo Dias de Novais was granted 140 miles of coastline with the right to conquer the interior, and he founded Luanda with 400 settlers in 1575. The wars of conquest launched by Novais lasted until 1680. No silver mines were found, but the Portuguese gained a precarious hold over the lower Kwanza River and parts of the Malanje highlands, and they occupied Benguela in 1617. The Dutch, French, and British undermined the Portuguese trade monopoly in Angola from the 17th century. The Dutch even seized Luanda and Benguela in 1641, but they were driven out by an expedition sent from Brazil in 1648.

The slave trade dominated Angola's economy, although some wax, ivory, and copper were also exported. Portuguese sugar plantations, first in São Tomé and then in Brazil, used slaves, and about a third of all slaves exported from Africa across the Atlantic left from the ports of Angola and Congo. Angola was not densely populated, and the slaving frontier expanded continuously, reaching the centre of Africa in the 18th century. The Portuguese raided for some of the slaves they exported, but from the end of the 17th century they usually bought them from Africans, who procured slaves through raiding, debt, judicial procedures, and the sale of children in times of famine. The introduction of corn (maize) and cassava from the Americas may to some extent have counterbalanced the negative demographic effects of the slave trade.

Decline of Kongo

The expansion of the slave trade destroyed some kingdoms but raised up others. Kongo entered into steep decline in the 17th century, while the Loango kingdom flourished north of the Congo estuary. The Ndongo kingdom in the Malanje highlands rose with the slave trade, but it was destroyed when the Portuguese pushed inland in the 17th century, to be replaced by the Kasanje kingdom in the Kwango valley. Farther east, the Lunda empire emerged as a great slave supplier in the 18th century. In the Bié Plateau, half a dozen warrior kingdoms fought for supremacy. Southern Angola was little involved in the slave trade, for its population was small and its inhabitants could sell cattle to procure imports.

Colonial transition, 1820s–1910. The export of slaves was banned in Angola in 1836, but it was not until the Brazilian market was closed in the early 1850s that slave exports fell. Slavery itself was legally abolished in the Portuguese empire in 1875, but it continued in thinly disguised forms until 1911. Slaves were exported to the coffee and cocoa plantations of São Tomé from the 1860s and were used in Angola to produce coffee, cotton, sugar, and fish. But from the 1850s, exports were dominated by products hunted or collected by Africans, first ivory and wax and later wild rubber. The Ovimbundu turned from slave raiding to long-distance trade, and their caravans penetrated as far east as Lake Tanganyika. The Chokwe were expert hunters of elephants and collectors of wax and rubber, and they used their accumulated firearms to overthrow the Lunda empire in the 1880s. The Kasanje kingdom collapsed when illicit slave trading undermined the king's central slave mart.

Angolans closer to the coast were more affected by the slow expansion of Portuguese colonialism and by the loss of land to settlers. Cotton and sugar were grown from the 1840s in oasis plantations in the coastal strip, and immigrants from the Algarve built up the fishing industry. Spontaneously occurring stands of coffee drew the Portuguese to carve out plantations in the Malanje highlands from the 1830s, and work on the railway from Luanda to Malanje commenced in 1885. In 1902 the building of the Benguela Railway began, after a brief campaign against the Ovimbundu, to serve the Katanga (Shaba) mines in the Belgian Congo. Portuguese small farmers were settled in the Huíla highlands from the 1880s, to counterbalance an influx of Boer trekkers from South Africa, and the southern railway was begun in 1905. In the Maiombe forests of the far north, plantations of cacao and oil palms were laid out in the 1900s.

Angola's frontiers were finally determined by negotiations in Europe in 1891, but the Portuguese only administered areas with plantations and railways and introduced systematic taxation of Africans only in 1906. Many positions in the colonial administration were held by Angolan Creoles, who were accepted as full Portuguese citizens. The spread of British and American Protestant missions from the 1870s was countered by subsidizing French Roman Catholic missions.

Development of colonial adminstration

From colonial conquest to independence, 1910–75. The proclamation of the Republic of Portugal in Lisbon in late 1910, followed in 1926 by the creation of the authoritarian New State (Estado Novo), marked the advent of modern Portuguese colonialism. The authorities stamped out slavery and undertook the systematic conquest of Angola. By 1920, all but the remote southeast of the colony was firmly under Portuguese control. Kingdoms were abolished, and the Portuguese worked directly through chiefs, headmen, and African policemen. Conversions to Christianity increased, and by 1940 there were about a million Christians in Angola, some three-quarters of them Roman Catholics. Angolan "natives" were taxed and subjected to forced labour and forced cultivation. A stringent set of tests were imposed on the few nonwhite "assimilated persons" who applied to be exempted from these impositions. Creoles were increasingly replaced by Portuguese immigrants in the administration.

Angola's economy was modernized and bound to that of Portugal by a system of protective tariffs. A network of dirt roads was built, and the Benguela Railway was completed to the boundary of the Belgian Congo in 1928. Lorries and fixed stores replaced trading caravans. Coffee, sugar, palm products, and sisal came mainly from the estate sector, and corn and cattle from smallholders. The cultivation of raw cotton for Portuguese textile mills was imposed by force. Alluvial diamond mining dominated the northeast from 1912, the fishing industry expanded, and a start was made with import-substituting industries.

After the independence of the Belgian Congo (later Zaire) in 1960, a major revolt rocked northern Angola in 1961; it was followed by a long guerrilla war. Land alienation and forced labour sparked off the rebellion in the coffee zone, while in the Kwango valley the peasants rose against forced cotton cultivation. In Luanda, an attack on the prison was led by frustrated Creoles. To contain the revolt, the Portuguese deployed large numbers of troops, set up strategic hamlets (forced settlements of rural Angolans), and, by encouraging Portuguese peasants to immigrate to Angola, raised the white population to about 330,000 by 1974. At the same time, they tried to "win hearts and minds" by abolishing forced cultivation, forced labour, and the stringent tests to gain "assimilated" status. They also stepped up the provision of education, health, and social welfare services and protected peasants from land alienation. The economy entered into a period of sustained boom, marked by rapid industrialization and the growth of oil production, and urban workers and many rural producers enjoyed rising standards of living.

The armed struggle continued, but the anticolonial guerrillas were seriously weakened by dissension. The Popular Liberation Movement of Angola (Movimento Popular de Libertação de Angola; MPLA) was founded in 1956 with the help of the clandestine Portuguese Communist Party; it was led by Agostinho Neto from 1962. It was popular in Luanda and among some rural Mbundu, and it drew foreign support from the Soviet Union. Initially based in Brazzaville, the MPLA moved to Zambia in 1965. The National Front for the Liberation of Angola (Frente Nacional de Libertação de Angola; FNLA), founded in 1957 under another name and led by Holden Roberto, drew its support from the Bakongo and some rural Mbundu. Based in Zaire, the FNLA obtained aid from the United States and China. In 1966 Jonas Savimbi set up a third movement, the National Union for the Total Independence of Angola (União Nacional para a Independência Total de Angola; UNITA), with a predominantly Ovimbundu leadership and with some support from the Chokwe and Ovambo. UNITA enjoyed little foreign backing (although China provided some aid) and lacked a secure foreign base,

Founding of MPLA, FNLA, and UNITA

since Zambia leaned toward the MPLA. The divisions between and within these three movements, which at times degenerated into armed conflict, allowed the Portuguese to gain the upper hand by the early 1970s. When a military coup in Portugal overthrew that country's dictatorship in April 1974, all three guerrilla movements had been almost entirely expelled from Angolan soil.

Independence. The three liberation movements proved unable to constitute a united front after the Portuguese coup. The FNLA's internal support had dwindled to a few Kongo groups, but it had strong links with the Zairean regime and was well armed; it thus made a bid to seize Luanda by force. The MPLA, with growing backing from the Portuguese Communist Party, Cuba, and the Soviet Union, defeated this onslaught and then turned on UNITA, chasing its representatives out of Luanda. UNITA was militarily the weakest movement, but it had the greatest potential electoral support, given the predominance of the Ovimbundu within the population, and it thus held out most strongly for elections. But the Portuguese army was tired of war and refused to impose peace and supervise elections. The Portuguese therefore withdrew from Angola in November 1975 without formally handing power to any movement, and nearly all the white settlers fled the country.

Portuguese withdrawal

The MPLA, in control of the capital city, declared itself the government of independent Angola and managed to win recognition from many African countries. UNITA and the FNLA set up a rival government in Huambo and called on a South African column to eject the MPLA from Luanda. The Cubans poured in troops to defend the MPLA, pushed the internationally isolated South Africans out of Angola, and gained control of all the provincial capitals. The Cuban expeditionary force, which eventually numbered some 40,000 to 50,000 soldiers, remained in Angola to pacify the country and ward off South African attacks. In 1977 the MPLA crushed an attempted coup by one of its leaders and, after a thorough purge, turned itself officially into a Marxist-Leninist party. The transformation of the economy along communist lines was pursued, with disastrous results. The major exception was the oil industry, which, managed by foreign companies, grew rapidly enough to enable Angola to stave off economic and military collapse. President Neto died in 1979 and was succeeded by the former minister of planning, José Eduardo dos Santos.

The FNLA withered away in exile, but UNITA reorganized itself with foreign backing—notably from the United States and South Africa—as an effective guerrilla force. Failed campaigns launched by the MPLA in 1987–88 and the increasingly effective UNITA attacks on oil installations forced the MPLA to adopt a more conciliatory posture, but a historic meeting between dos Santos and Savimbi in June 1989 produced only a short-lived ceasefire.

With the collapse of communism in eastern Europe, in mid-1990 the MPLA abandoned the one-party state and produced a new constitution allowing free elections that included UNITA. Cuban troops withdrew in 1991, and, although an agreement to disarm all forces had not yet been met, elections were held in 1992 under United Nations supervision. Dos Santos was elected president, and the MPLA gained a majority in the parliament. UNITA made a strong showing but, charging election fraud, renewed the civil war. At the end of 1992, UNITA controlled approximately two-thirds of the country; this area included Angola's wealthiest diamond mines, which funded UNITA's operations. Fighting raged throughout 1993 as the government gradually regained territory, and international pressure mounted for the two sides to reach a peaceful solution. Finally, the combatants signed an agreement called the Lusaka Accord, which allowed UNITA to be reintegrated into the government. Although minor fighting between the two groups continued, dos Santos and Savimbi met several times over the next three years to resolve issues relating to the final form of the combined government. UNITA delegates formally joined the government in 1997, but relations between the two groups were further complicated that year by the civil war in Congo (Kinshasa). UNITA supported the crumbling government regime because it had been a conduit for UNITA's illicit diamond trade, while the Angolan government supported the rebels led by Laurent Kabila.

By the end of the decade, hostilities between the government and UNITA had resumed, and the UNITA delegates had been expelled from the government. The issue was not resolved until February 2002 when government troops caught up with Savimbi and killed him. A cease-fire between UNITA and the government was signed in April, and the war was officially declared over in September; the government focused on rebuilding the crumbling infrastructure and stabilizing the economy. (W.G.C.-S./Ed.)

For later developments in the history of Angola, see the BRITANNICA BOOK OF THE YEAR.

Botswana

The Republic of Botswana is a landlocked country with an area of 224,607 square miles (581,730 square kilometres) in the centre of southern Africa. The capital is Gaborone. The territory is roughly square—approximately 600 miles from north to south and 600 miles from east to west—with its eastern side protruding into a sharp nose. Its eastern and southern borders are marked by river courses and an old wagon road; its western borders are lines of longitude and latitude through the Kalahari, and its northern borders combine straight lines with a river course.

Botswana is bounded by Namibia to the west and north (the Caprivi Strip), Zambia and Zimbabwe to the northeast, and South Africa to the southeast and south. The Zambezi River border with Zambia is only several hundred yards long. The border along the main channel of the Chobe River up to the Zambezi is disputed with Namibia. The point at which the borders of Botswana, Namibia, Zambia, and Zimbabwe meet in the middle of the river has therefore never been precisely determined.

Before its independence in 1966, Botswana was a British protectorate known as Bechuanaland. It was also one of the poorest and least-developed states in the world. The country is named after its dominant ethnic group, the Tswana, or Batswana ("Bechuana" in older variant orthography). The national language is Setswana (or Sechuana), and the official language is English.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief and drainage.* Botswana extends from the Chobe River (which drains through the Zambezi to the Indian Ocean) in the north to the Molopo River (part of the Orange River system, which flows into the Atlantic) in the south. To the east it is bordered by the Limpopo River and its tributaries, the Ngotwane (Notwani), Marico (Madikwe), and Shashe.

The country has a mean altitude of 3,300 feet and consists largely of a sand-filled basin, with gently undulating plains rising to highlands in neighbouring countries. The highest point is 4,888 feet (1,490 metres) in the hills north of Lobatse in southeastern Botswana; the lowest point is 2,170 feet at the country's easternmost point, in the Limpopo valley.

The country is divided into three main environmental regions. The hardveld region consists of rocky hill ranges and areas of shallow sand cover in eastern Botswana. The sandveld region is the area of deep Kalahari sand covering the rest of the country. The third region consists of ancient lake beds superimposed on the northern sandveld in the lowest part of the Kalahari Basin.

Geologic exploration has been limited by the depth and extent of Kalahari sand covering the surface geology. The rock groups underlying most of the sandveld are therefore the least-known but appear to be the youngest, belonging to the Karroo (Karoo) System, formed 290 to 208 million years ago. Elsewhere, Precambrian rock formations predominate. The surface geology of the eastern hardveld, exposed in its hill ranges, largely consists of basement complex rocks (more than 2.5 billion years old) intruding from the Transvaal and southern Zimbabwe. This complex is known to extend into younger rock formations (2.5 to 1.2 billion years old) in the extreme southern sandveld,

Geology

while rocks of the Ghanzi and Damara groups (1.2 billion to 570 million years old) extend across the northwest corner of the country into northern Namibia.

Drainage through the marshes of the Okavango delta is complex and imperfectly understood. The perennial Okavango River runs southward into its delta across the Caprivi Strip from the highlands of Angola. Most of its water evaporates from the 4,000 square miles of the delta wetlands. Floodwater reaches down through the eastern side of the marshes to the Boteti River, which flows sporadically to Lake Xau (Dow) and the Makgadikgadi Pans (also roughly 4,000 square miles in area). Less and less water has been flowing through the western side of the Okavango marshes during the 20th century, so that 70-square-mile Lake Ngami—famous a century ago—is today dry and almost unrecognizable as a lake. Meanwhile, the eastern Makgadikgadi Pans are annually flooded by the otherwise ephemeral Nata River from the Zimbabwe highlands, while the southern tributaries of the pans are now dry fossil valleys.

Islands of papyrus and other aquatic vegetation in the delta of the Okavango River, northern Botswana.

The Molopo River and its Ramatlhabama tributary, on the southern border of Botswana with a course flowing into the Orange River, today rarely flood more than 50 miles from their sources. Most rivers in Botswana are ephemeral channels, usually not flowing above ground except in the summer rainy season. The two great exceptions to this rule are vigorous channels fed by the rains of central Africa—the Okavango River above its delta and the Chobe River flowing through its marshes along the northern border to join the Zambezi above the Victoria Falls.

Soils. The soils of the eastern hardveld consist of moderately dry red loamy *mokata* soils on the plains, or mixed chalky and sandy *chawana* soils, with brownish rocky *seloko* soils on and around hills. *Seloko* soils are considered best for grain crops. The fertility of all soils is limited by the amount of rainfall, which is sometimes inadequate on the hardveld and regularly unable to support any cultivation on the sandveld.

Soil fertility

The alluvial soils of the ancient lake beds include gray loamy soils in the wetlands, gray-green saline soils on the pans, gray clayish soils to yellowish sandy soils around the wetlands, and very chalky light gray soils around the pans. There are also areas of gray to black cracking clay in former wet areas, such as those around Pandamatenga.

Climate. The annual climate ranges from months of dry temperate weather during winter to humid subtropical weather interspersed with drier periods of hot weather during summer. In summer (which lasts from October to March) temperatures rise to about 93° F (34° C) in the extreme north and southwest, the warmest parts of the country. In winter (which lasts from April to September), there is frequent frost at night, and temperatures may fall to near freezing in some high-altitude areas during the day. Summer is heralded by a windy season, carrying dust from the Kalahari, from about late August to early October. Annual rainfall, brought by winds from the Indian Ocean, averages 18 inches (460 millimetres), representing a range from 25 inches in the extreme northeast to less than 5 inches in the extreme southwest. The rains are almost entirely limited to summer downpours between December and April, which also mark the season for plowing and planting. Cyclic droughts, often lasting up to five or six years in every two decades, can limit or eliminate harvests and reduce livestock to starvation.

Plant and animal life. The Kalahari sandveld has often been called "thirstland" to distinguish it from true desert. Even in its southwestern corner, where there are some bare sand dunes, the vegetation is more characteristic of dry steppe than desert.

The general vegetation of the country is savanna grassland with yellow or light brown grass cover (turning green after rains) and woody plants. The savanna ranges from acacia shrub savanna in the southwest through acacia thornbush and tree savanna "parkland" into denser woodland and eventually forest as one moves north and east. Croton and Combretum tree savanna is found on the rocky hills of the eastern hardveld. Acacia tree savanna merges northward into mopane (African ironwood) savanna woodland. Mopane woodland covers most of the northern and eastern third of the country, with the exception of the open grasslands immediately surrounding the Okavango delta and Makgadikgadi Pans.

Animal life is extremely varied in a thirstland environment. About 150 species of mammals are found in Botswana. These range from 30 species of bats and 27 of rodents to more than 30 species of large mammals. Birdlife is prolific, with more than 460 species.

Botswana has a great variety of reptiles and amphibians, of which more than 200 species have been described in detail. The principal fish, in the rivers of the north, are tilapia (African bream), catfish, and the tigerfish, which is famous for its ferocious resistance to being caught on a line.

Settlement patterns. The human and livestock population of Botswana is concentrated around the hill ranges of the eastern hardveld and along the perennial rivers of the north. Approximately one-third of the population lives in scattered rural settlements, usually based around livestock pens. More than half the population lives in rural village settlements of more than 1,000 people, including traditional towns with up to 40,000 people.

Village patterns

The typical rural settlement and land use pattern of the eastern hardveld in the past may be characterized as having been concentric circles around a concentrated village nucleus. The family had a home base in the village, where the majority of its members spent most of the year. In the appropriate season it cultivated lands (fields) within one or two days' walk from the village. The family cattle, on the other hand, were pastured for most of the year at "cattle-posts" a number of days' walk from the village. Finally, beyond the cattle-posts there were hunting lands.

The villages and traditional towns of Botswana are still basically laid out around the *kgotla* (courtyard) and cattle kraal (corral) of traditional rulers and are subdivided into wards, each of which mimics the village or town plan with its own central *kgotla* and kraal. But, especially since the 1970s, traditional settlements have been sliced through by modern roads and facilities such as schools and offices, as well as shopping malls and bars. Traditional architecture of thatch roofing and clay walls has given way to corrugated metal roofing and brick walls.

Two of the seven larger towns of Botswana, Francistown (1897) and Lobatse (1902), originated as small urban centres on the railway for white farming communities. Both

began to develop in size and function in the 1950s, as employment in nonagricultural services expanded. Gaborone, the capital city, was founded in 1964; since then its population has grown to more than 150,000. Selebi-Phikwe (1971) and Jwaneng (1979) constitute the only substantial mining towns; the smaller diamond town of Orapa (1971) is under company control behind high security fences. The newest mining town, Sua (1991), is based on the soda-ash deposits of the eastern Makgadikgadi Pans.

The people. *Ethnic groups.* The dominant ethnic identity in Botswana is Tswana. The country's whole population is characterized as Batswana (singular, Motswana) whatever their ethnic origin. Though no attempt to count population by ethnic origin has been made since 1946, probably less than half the population is "ethnic Tswana" by origin. There are far greater numbers of ethnic Tswana in South Africa.

The Tswana

Tswana ethnic dominance ("Tswanadom") in Botswana can be dated to the eight Tswana states which ruled most of the area in the 19th century. Under British colonial rule, the populations of these states were given the official status of "tribes," a term still used officially today.

Within southeastern Botswana the other main ethnic identity besides Tswana, that of the Khalagari (Western Sotho), has become so incorporated as to be almost indistinguishable from the Tswana. Even their name is now usually rendered in the Tswana form as "Kgalagadi."

The Ngwato of east-central Botswana constitute the largest traditional "tribal" state but are probably less than one-fifth ethnic Tswana by origin. The major incorporated ethnic groups are Khalagari, Tswapong and Birwa (both Northern Sotho), and Kalanga (Western Shona). With larger numbers to the east in Zimbabwe, some Kalanga have resisted full incorporation.

The Tawana state of northwestern Botswana can be seen as the least successful in incorporating other ethnic groups. Most of its population is Yei and Mbukushu by origin, related to riverine peoples in the Caprivi Strip, Angola, and Zambia to the north. Smaller numbers of Mbanderu and Herero have greater numbers of close relatives across the border in Namibia. The Subiya along the Chobe, closely related to people in the Caprivi Strip and Zambia, were excluded from the Tawana "tribal" reserve by the British.

Small scattered groups of Khoisan people inhabit the southwestern districts of Botswana, as well as being incorporated with other ethnic groups. The Khoisan speak languages characterized as Khoe, or Khwe, and San. They include communities with their own headmen and livestock, as well as poorer groups employed by Tswana and white cattle farmers.

White settlement in Botswana, consisting of some Afrikaners and fewer English settled in border farms, totaled fewer than 3,000 people in the colonial period.

Religious groups. Christianity, brought by missionaries from the south such as David Livingstone, was established as the official religion of the eight Tswana states by the end of the 19th century. Indigenous religious and medical practices, notably respect for patriarchal ancestors, either declined or were assimilated within popular Christian beliefs. Allegiance to the old state churches, notably those of the Congregationalists (London Missionary Society), has declined since the 1950s. The two most active and popular churches are the Zion Christian church (based in South Africa) among the working class and the Roman Catholic church among the middle class. There are also numerous other small Zionist and Apostolic churches in rural villages, as well as United Reformed (Congregational and Methodist), Dutch Reformed, and Anglican churches, and predominantly expatriate Muslim, Quaker, Hindu, and Bahā'ī congregations.

Demographic trends. After six previous censuses of variable quality, Botswana had its first systematic national census in 1964. Total population was estimated at 550,000, with 35,000 absentees—mostly adult male workers in South Africa. Since 1964, the population has grown at about 3.4 percent a year, thus exceeding 1,000,000 in the early 1980s and doubling every 20 years. Meanwhile, the rate of labour migration abroad has been reduced by a combination of restrictions by South Africa and increased employment opportunities at home. Between the 1950s and the 1970s Botswana also provided a home and eventual citizenship for significant numbers of refugees from South Africa, Angola, and Zimbabwe.

Population growth

The age and gender composition of the country is weighted by an increasingly youthful population: approximately one-fifth are under age 5, and nearly half are younger than age 15. Females exceed males in age groups over 15; below that age the gender ratio is more or less equal because of recently reduced male infant mortality. The 1991 census showed improved life expectancy of 63.1 years at birth for females and 57 years for males, over the 1981 figures of 61.2 and 54.7 years, respectively. Botswana is also the first African country to experience a falling birth rate in response to improved life expectancy.

The economy. Botswana has a free market economy with a strong tradition of central government planning to provide infrastructure for private investment. The economy has grown rapidly since the mid-1960s, with the per capita gross domestic product increasing from less than $50 to more than $1,000 by the mid-1980s.

Less than a quarter of the adult work force is in formal paid employment. Relatively few rural households benefit from cattle sales: almost half of them have no cattle, and less than 10 percent own about half of the country's cattle (averaging 100 head each). Few households produce enough crops to cover even their own subsistence, let alone to sell on the market. Four out of five rural households survive on the income of a family member in town or abroad. That still leaves a significant number of rural households, usually female-headed, with no source of income known to statisticians.

State revenues reaped from mining development have been spent on basic rural infrastructure and welfare services and on schemes to subsidize the development of cattle and crop production, which have in general benefited the richer rural households. Trade unionism, restricted by legislation, is as yet underdeveloped in Botswana.

Resources. Diamonds, the major economic resource of the country, have been exploited on a large scale since 1970. They are mined from some of the world's largest diamond pipes at Orapa and Letlhakane, south of the Makgadikgadi Pans, and at Jwaneng in the southeastern sandveld. Nickel and copper have been mined at Selebi and Phikwe near the Motloutse River since 1974. Coal is mined for power generation at Morupule near Palapye. Botswana's other major proven mineral resources are salt and soda ash, which was fully exploited at Sua on the eastern Makgadikgadi Pans from 1991.

Diamonds

Surface water resources are limited to the wetlands and perennial rivers in the north and three major dam lakes at Gaborone, Shashe, and Mopipi (serving Orapa). Plans are under active consideration to canalize water through the Okavango wetlands toward Mopipi via a holding dam at Maun. Underground water is tapped in large quantities near Palapye and south of Gaborone.

Agriculture. Most of the population is partially engaged in agricultural production, but there is little land suitable for productive cultivation. Agricultural output constitutes less than one-tenth of gross national product, and most of that is in the form of livestock production for urban and export markets. Grain production (mostly sorghum and corn) has fallen short of national consumption for most of the 20th century, and foodstuffs from South Africa and Zimbabwe are Botswana's major import commodities. Fishing and forestry production are limited and largely confined to the extreme north.

Botswana, with terrain comparable to Texas or Australia, is traditionally seen as cattle country. Given sufficient water and pasture, and controls on the spread of hoof-and-mouth disease from wetland buffalo, it is a healthy environment for raising high-bulk, high-quality indigenous beef cattle. The government has invested heavily in disease prevention, modern abattoirs, and support services to cattle producers. Because of drought the national herd has fluctuated between one and three million head since the 1960s, with an export offtake of up to a quarter of a million per annum and a growing internal market. Various schemes—so far unsuccessful—have been attempted to

improve range management. Meanwhile, the main export market for beef, the European Community, has become increasingly unreliable. (Ne.P.)

Limits on development

Industry. Industrial development in Botswana has been limited by the high costs of power and water, the lack of appropriate management and labour skills, and the small domestic market. Manufacturing activity up to the 1980s largely consisted of meat processing at Lobatse in the south. In the early 1980s capital and textile production were transferred from Zimbabwe to nearby Francistown in Botswana, and diamond sorting and service industries grew in the booming capital city, Gaborone.

The national electric power grid, serving mines and eastern towns, is based on a large coal-powered generating station at Morupule near Palapye, supplemented by connections to the Zimbabwean and South African national grids.

Tourists are attracted to Botswana by relatively unpopulated and "remote" wetland and thirstland environments. Government policy is to limit the density and environmental impact of tourism through licensing of a limited number of high-cost safari companies.

Finance. The Botswana economy is regulated by a central bank, the Bank of Botswana, and a strong ministry of finance and development planning. There are two major multinational commercial banks, with branch operations that extend to village level. Botswana has had the unusual problems, for a developing country, of a government budget surplus running into billions of dollars and excess capital lying unutilized in private banks.

The budget surplus and bank liquidity were partially depleted by diversion into a construction boom in the late 1980s and early '90s, including infrastructure for new mining operations and military airports. A small stock exchange has been set up. The economy, from diamonds to nickel-copper to soda ash and construction, remains dominated by South Africa's De Beers–Anglo American Corporation conglomerate.

Trade. Botswana sends the great bulk of its exports to the world market beyond Africa, mainly to Europe and North America. It takes nearly three-fourths of its imports from South Africa, its neighbour with a gross national product vastly larger than that of Botswana. Imports consist of machinery and transport equipment, food products, and consumer goods, often manufactured or serviced by multinational companies based in South Africa. From Zimbabwe, Botswana imports mainly food products. Other imports from the rest of the world consist largely of high-technology equipment.

Botswana is joined in a customs union with South Africa, Lesotho, Swaziland, and Namibia (the Southern African Customs Union) and, along with the other countries of southern Africa, is a member of the Southern African Development Community (SADC). Domestic trade patterns within Botswana are dominated by large, mostly foreign-owned wholesale operations and large foreign retailers in urban areas, though there is also an increasing proliferation of small stores owned by citizens.

Transport. The 400-mile railway along the eastern side of the country was completed in 1897, linking South Africa and Zimbabwe, but had limited impact on the Botswana economy until the 1970s, when the first branch lines were opened to serve mining areas. At independence in 1966, there were only a few miles of paved roads—all inside town boundaries. Since then, the major towns have been linked by paved main highways. Most of the sandveld, however, is accessible only to four-wheel-drive vehicles.

International air traffic in Botswana, though dating to 1919, was limited until the opening up of the Sir Seretse Khama Airport at Gaborone in 1984. Gaborone is now served by British and French airlines as well as by regional airlines and the national parastatal airline, Air Botswana.

Administration and social conditions. *Government.* Botswana is a unitary state with a multiparty parliamentary system and an executive presidency. Since independence, Botswana has held free elections every five years, a relatively uncorrupt bureaucracy, and judicial respect for human rights and the rule of law. The government has also distributed increasing resources widely if not always equally among the people.

The National Assembly

Parliament consists of a National Assembly of elected members (elected by universal adult suffrage in single-member constituencies) and a handful of ex officio members nominated by the ruling political party. There is also a House of Chiefs, with an advisory role on matters of legislation pertaining to tribal law and custom.

The ruling party, first elected in 1965 and reelected at five-year intervals since then, is the Botswana Democratic Party. Its overwhelming majorities in elections have been based on rural support; opposition parties have drawn their strength generally from urban areas. The Botswana People's Party was the main opposition in the 1960s, when urban areas were small. Since then the Botswana National Front has grown in strength, holding the majority of the city council of Gaborone and both of Gaborone's parliamentary seats.

Local councils, rural and urban, have been elected since 1969 simultaneously with national parliamentary elections. The power of local councils is limited by the right of the central government to nominate ex officio voting members and by central government appointment of supervisory district commissioners and planning staff.

The evolving political and economic alignments of Botswana's foreign policy are indicated by the countries to which it has sent resident ambassadors—originally the United States, the United Kingdom, and Zambia in the 1960s, followed by Belgium, Zimbabwe, and Sweden in the 1980s and by Namibia, Russia, and China in the early 1990s. Full diplomatic relations were established with South Africa in 1994.

Education. Since independence, enrollment at all levels of education has increased steadily. Enrollments in primary education are still lower in the remote western and northwestern districts than in other areas, however, as poorer non-Tswana children often miss out on school.

International interest has been aroused by an alternative system of education, integrating vocational skills into the secondary curriculum, developed by the educationist Patrick van Rensburg at Swaneng Hill near Serowe. But "education-with-development" has had little impact on the general curriculum within Botswana's schools.

A university campus in Gaborone, founded in 1976, became the University of Botswana in 1982. Officially, more than three-fourths of the population is considered literate, although this is probably an overestimate. Rural literacy rates are higher in the east and northeast and lower in the west and northwest. More women than men are literate, as many boys traditionally are employed in cattle herding rather than being sent to school.

Health and welfare. Botswana has a dry and warm climate generally conducive to good health. The incidence of tropical diseases—notably malaria, bilharzia (a parasitic disease), and sleeping sickness—is limited by the environment and lack of surface water. The most common fatal diseases are intestinal (diarrheal and digestive diseases) and respiratory (pneumonia and tuberculosis).

The main threats to health are diseases associated with changing lifestyle, particularly diet. There has been increased incidence of high blood pressure, strokes, and heart disease, as well as dental caries in older children. The spread of AIDS has had a devastating effect in Botswana, where the rate of infection has been one of the highest in the world; by 2000 more than one-third of the adult population was infected with HIV, and the growing number of AIDS orphans loomed as a serious social problem.

Health policy

Since 1973, government health policy has been based on the provision of basic health services in the form of health posts in every village with a population of more than 500 and clinics in every area with more than 4,000 in a nine-mile radius. Since the late 1980s there has also been extensive investment in two large national referral hospitals, at Gaborone and Francistown.

The use of government health services is free of charge. There are also a number of Western-certified physicians in private practice and many traditional herbalists, healers, and diviners. (Ne.P./Ed.)

Cultural life. The cultural life of Botswana reflects the

dual heritage and intermingling of Tswana and English cultural domination. The two languages and cultures are subtly mixed and alternated in urban and official situations.

Western dress has been general among people in Botswana, except at the poorest level, since the late 19th century. Common diet and cuisine consist of sorghum and corn porridge, beans and pulses and traditional spinach, supplemented by tomato, potato, onion, and cabbage usually purchased from stores. Meat consumption has become more common with the opening of small butcheries selling beef. Traditional foods include dried *phane* caterpillars from mopane woodland, eaten as relish or snacks, fruits such as the wild *morula* plum, and beer made from sorghum or millet.

Families in rural villages live in traditional compounds, usually with two or three small houses of cylindrical clay walls and conical thatch roofs, set around an open fireplace and surrounded by low clay walls. Most recent houses are square with metal roofs, while many houses in the northwest are made of reed.

Rites of burial, marriage, and birth have been adapted to Christianity but remain extremely important in Botswana life.

Traditional arts

Traditional music, based on stringed instruments, and dance generally declined during the colonial period. After independence there was a revival of interest, particularly in music on the radio.

The best-known modern art form incorporating traditional craftwork is basketry—most of it from northwestern Botswana—which is widely exported overseas.

The author Bessie Head (1937–86) wrote novels in English that reflect the contemporary realities and history of Serowe. The publishing of fiction in Setswana was revived in the 1980s.

There is a national museum and art gallery in Gaborone and an increasing number of district museums founded by local community initiative. A national learned and scientific society, the Botswana Society, holds regular lectures and publishes an annual journal and books.

Football (soccer) is the national sport, played on fields and in stadiums across the country every Saturday.

The government issues a free daily newspaper, mostly in English, and runs a radio service, mostly in Setswana. There are also several separate private weekly newspapers, with circulation in eastern towns, and private local television stations, mostly relaying broadcasts from neighbouring countries. There is no government censorship. During the 1980s three multinational publishers set up branches to generate published materials for schools.

For statistical data on the land and people of Botswana, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The history of Botswana is in general the history of the Kalahari area, intermediate between the more populated savanna of the north and east and the less populated steppe of the south and west. Although reduced to a peripheral role in southern Africa for most of the 20th century, at other times Botswana has been a central area of historical development.

Early pastoral and farming peoples. *Khoisan-speaking hunters and herders.* People speaking Khoisan (Khoe and San) languages have lived in Botswana for many thousands of years. Depression Shelter in the Tsodilo Hills has evidence of continuous Khoisan occupation from about 17,000 BC to about AD 1650. During the final centuries of the last millennium before Christ, some of the Khoi (Tshu-khwe) people of northern Botswana converted to pastoralism, herding their cattle and sheep on the rich pastures revealed by the retreating lakes and wetlands.

Bantu-speaking farmers. Meanwhile, the farming of grain crops, and the speaking of Bantu languages, were carried gradually southward from the Equator. By about 20 BC such farmers were making and using iron tools on the upper Zambezi.

The earliest dated Iron Age site in Botswana is an iron-smelting furnace in the Tswapong Hills near Palapye, dated about AD 190 and probably associated with Iron Age farmers from the Limpopo valley. The remains of small beehive-shaped houses made of grass matting, occupied by early Iron Age farmers around Molepolole, have been dated from about AD 420. There is also evidence of early farming settlement west of the Okavango delta, in the Tsodilo Hills alongside Khoisan hunter and pastoralist sites, dated from about AD 550. Archaeologists therefore have difficulty in interpreting the hundreds of rock paintings in the Tsodilo Hills, which were once assumed to be painted by "Bushman" (San) hunters remote from all pastoralist and farmer contact.

Iron Age states and chiefdoms. *Eastern states and chiefdoms.* From about AD 1095 southeastern Botswana saw the rise of a new culture, characterized by a site on Moritsane hill near Gabane. The Moritsane culture is historically associated with the Khalagari (Kgalagadi) chiefdoms, the westernmost dialect group of Sotho (or Sotho-Tswana) speakers.

The Toutswe state

The area within 50 or 60 miles of Serowe saw a thriving farming culture, dominated by rulers living on Toutswe hill, between about the 7th and 13th centuries. The prosperity of the state was based on cattle herding, with large corrals in the capital town and in scores of smaller hilltop villages. (Ancient cattle corrals are identified by the peculiar grass growing on them.) The Toutswe people also hunted westward into the Kalahari and traded eastward with the Limpopo.

The Toutswe state appears to have been conquered by its neighbour, the Mapungubwe state, centred on a hill at the Limpopo-Shashi confluence, in the 13th century. But Mapungubwe's triumph was short-lived, as it was superseded by the new state of Great Zimbabwe, north of the Limpopo River. Great Zimbabwe's successor from about 1450, the Butua state based at Khami (Kame) near Bulawayo in western Zimbabwe, controlled trade in salt and hunting dogs from the eastern Makgadikgadi Pans, around which it built stone-walled command posts.

Western chiefdoms. From about AD 850 farmers from the upper Zambezi, ancestors of the Mbukushu and Yei peoples, reached as far south and west as the Tsodilo Hills (Nqoma). The oral traditions of Herero and Mbanderu pastoralists, west of the Okavango, relate how they were split apart from their Mbandu parent stock by 17th-century Tswana cattle-raiding from the south.

Rise of Tswanadom. During the 13th and 14th centuries AD a number of powerful dynasties began to emerge among the Tswana in the western Transvaal region. Rolong chiefdoms spread westward over lands controlled by Khalagari peoples. Khalagari chiefdoms either accepted Rolong rulers or moved westward across the Kalahari.

The main Tswana dynasties of the Hurutshe, Kwena, and Kgatla were derived from the Phofu dynasty, which broke up in its home in the western Transvaal region in the 16th century. The archaeology of the Transvaal region shows that, after about 1700, stone-walled villages and some large towns developed on hills. These states were probably competing for cattle wealth and subject populations, for control of hunting and mineral tribute, and for control of trade with the east coast.

The Ngwaketse chiefdom

Growth of Tswana states. Kwena and Hurutshe migrants founded the Ngwaketse chiefdom among the Khalagari-Rolong in southeastern Botswana by 1795. After 1750 this chiefdom grew into a powerful military state controlling Kalahari hunting and cattle raiding and copper production west of Kanye. Meanwhile, other Kwena had settled around Molepolole, and a group of those Kwena thenceforth called Ngwato settled farther north at Shoshong. By about 1795 a group of Ngwato, called the Tawana, had even founded a state as far northwest as Lake Ngami.

Times of war. From about 1750, trading and raiding for ivory, cattle, and slaves spread inland from the coasts of Mozambique, the Cape Colony, and Angola. By 1800, raiders from the Cape had begun to attack the Ngwaketse. By 1824 the Ngwaketse were being attacked by the Kololo, a military nation on the move that had been expelled northwestward by raiders from the east. The great Ngwaketse warrior king Makaba II was killed, but the Kololo were pushed farther north by a counterattack in 1826.

The Kololo moved through Shoshong to the Boteti River, expelling the Tawana northward. In about 1835 the Kololo settled on the Chobe River, extending their power to the upper Zambezi, until their final defeat there by their Lozi subjects in 1864. The Kololo were followed by the Ndebele, a military nation led by Mzilikazi, who settled in the Butua area of western Zimbabwe in 1838–40, after the local Rozvi state was conquered.

Prosperous trading states. The Tswana states of the Ngwaketse, Kwena, Ngwato, and Tawana were reconstituted in the 1840s after the wars ended. The states competed with each other to benefit from the increasing trade in ivory and ostrich feathers being carried by wagons down new roads to the Cape Colony in the south. Those roads also brought Christian missionaries to Botswana and Boer trekkers who settled in the Transvaal to the east.

The most remarkable Tswana king of this period was Sechele (ruled 1829–92) of the Kwena around Molepolole. He allied himself with British traders and missionaries and was baptized by David Livingstone. He also fought the Boers, who tried to seize people who fled from the Transvaal to join Sechele's state. But by the later 1870s the Kwena had lost control of trade to the Ngwato under Khama III (ruled 1872–73; 1875–1923), whose power extended to the frontiers of the Tawana in the northwest, the Lozi in the north, and Ndebele in the northeast.

British protectorate. White miners and prospectors flooded Botswana in 1867–69 to start deep gold mining at Tati near Francistown. But the gold rush was short-lived, and the diamond mines at Kimberley south of Botswana became southern Africa's first great industrial area from 1871. Migrant labourers from Botswana and countries farther north streamed to Kimberley and later to the gold mines of the Transvaal.

The "Scramble for Africa" in the 1880s resulted in the German colonization of South West Africa. The new German colony threatened to join across the Kalahari with the independent Boer republic of the Transvaal. The British in the Cape Colony responded by using their missionary and trade connections with the Tswana states to keep the roads through Botswana open for British expansion to Zimbabwe and the Zambezi. In 1885 the British proclaimed a protectorate over their Tswana allies and the Kalahari as far north as the Ngwato; the protectorate was extended to the Tawana and the Chobe River in 1890.

Threats of incorporation

In 1890 British colonial expansion was privatized in the form of the British South Africa Company, which used the road through the Bechuanaland Protectorate to colonize what would soon be called Rhodesia. But the protectorate itself remained under the British crown, and white settlement remained restricted to a few border areas, after an attempt to hand it over to the company was foiled by a delegation of three Tswana kings to London in 1895. The kings, however, had to concede to the company the right to build a railway to Rhodesia through their lands.

The British government continued to regard the protectorate as a temporary expedient, until it could be handed over to Rhodesia or, after 1910, to the new Union of South Africa. Hence the administrative capital remained at Mafikeng—actually outside the protectorate's borders in South Africa—from 1895 until 1964. Investment and administrative development within the territory were kept to a minimum. It declined into a mere appendage of South Africa, for which it provided migrant labour and the rail transit route to Rhodesia. Short-lived attempts to reform administration and to initiate mining and agricultural development in the 1930s were hotly disputed by leading Tswana chiefs, on the grounds that they would only enhance colonial control and white settlement. The territory remained divided into eight largely self-administering "tribal" reserves and five white settler farm blocks, with the remainder classified as crown (*i.e.*, state) lands.

The extent of the Bechuanaland Protectorate's subordination to the interests of South Africa was revealed in 1950. In a case that caused political controversy in Britain and the empire, the British government barred Seretse Khama from the chieftainship of the Ngwato and exiled him from Botswana for six years. This, as secret documents have since confirmed, was in order to satisfy the South African government, which objected to Seretse Khama's marriage to a white Englishwoman at a time when racial segregation was being reinforced in South Africa under apartheid.

Advance to independence. From the late 1950s it became clear that Bechuanaland could no longer be handed over to South Africa and must be developed toward political and economic self-sufficiency. The supporters of Seretse Khama began to organize political movements from 1952, and there was a nationalist spirit even among older, more ethnic-based leaders. A legislative council was eventually set up in 1961 after limited national elections. The Bechuanaland People's Party was founded in 1960, and the Bechuanaland Democratic Party (BDP)—led by Seretse Khama—in 1962.

Founding of political parties

After long resistance to constitutional advance before economic development could pay for it, the British began to push political change in 1964. A new administrative capital was rapidly built at Gaborone. Bechuanaland became self-governing in 1965, under an elected BDP government with Seretse Khama as prime minister. In 1966 the country became the Republic of Botswana, with Seretse Khama as its first president.

For its first five years of political independence, Botswana remained financially dependent on Britain to cover the full cost of administration and development. The planning and execution of economic development took off in 1967–71 after the discovery of diamonds at Orapa. The essential precondition for this was renegotiation of the customs union with South Africa, so that state revenue would benefit from rising capital imports and mineral exports rather than remain at a fixed percentage of total customs union income. This renegotiation was achieved in 1969.

Botswana since independence. From 1969 Botswana began to play a more significant role in international politics, putting itself forward as a nonracial, liberal democratic alternative to South African apartheid. South Africa was obliged to step down from its objections to Botswana's building a road, with U.S. aid, to Zambia, avoiding the old railway and road route through Rhodesia. From 1974 Botswana was—together with Zambia and Tanzania and later Mozambique and Angola—one of the "Frontline States" seeking to bring majority rule to Rhodesia, Namibia, and South Africa.

Economic and political growth

With an economy growing between 12 and 13 percent annually, Botswana extended basic infrastructure for mining development and basic social services for its population. More diamond mines were opened, on relatively favourable terms of income to the state, and less economically successful nickel and copper mining commenced at Selebi-Phikwe. The BDP was consistently reelected with a large majority, though the Botswana National Front (BNF; founded 1965) became a significant threat after 1969, when traditional leaders joined the socialists in BNF ranks attacking the government's policies.

The late 1970s saw civil war in Rhodesia and urban insurrection in South Africa, from which refugees flooded into Botswana. Botswana played a part in the final settlement of the Rhodesian war, resulting in independence for Rhodesia (now called Zimbabwe) in 1980. Its main contribution, however, was in formulating the Southern African Development Community (SADC) to look to the future of the region. The idea behind SADC, as expounded by Seretse Khama, was to coordinate disparate economies rather than to create a unified market in southern Africa. All the states of southern Africa, except South Africa and Namibia, formed SADC in 1980, to work together in developing identified sectors of their economies, particularly the transport network to the ports of Mozambique.

Seretse Khama died in 1980 and was succeeded by Quett Masire, who had been his deputy since 1965. The economy continued to expand rapidly after a temporary slump in diamond and beef exports at the beginning of the 1980s. Meanwhile, the country experienced the problems of an increasing gap between urban rich and rural poor.

Between 1984 and 1990, Botswana suffered from upheavals in South Africa when South African troops raided the Frontline States. A new era in southern African relations has opened since the independence of Namibia in

1990 and political changes in South Africa in 1994. Masire retired from politics in March 1998 and was succeeded by his vice president, Festus Mogae, as required by the constitution. General elections held in 1999 kept the BDP in power. Botswana's major problem at the beginning of the 21st century was its large number of HIV/AIDS cases, estimated to include nearly two-fifths of the population. (Ne.P.)

For later developments in the history of Botswana, see the BRITANNICA BOOK OF THE YEAR.

Lesotho

The Kingdom of Lesotho in southern Africa has an area of 11,720 square miles (30,355 square kilometres). Like only two other independent states in the world (Vatican City and the Republic of San Marino), Lesotho is completely encircled by a single country on which it must depend exclusively for access to the outside world. It forms an enclave within the Republic of South Africa, bordering on three of the latter's provinces—KwaZulu/Natal, Free State, and Eastern. This physical dependence on its neighbour is further accentuated by Lesotho's own grave lack of resources, which makes it necessary for one-fourth to one-half of the labour force to live and work, mainly as migrants, in South Africa. For this reason, Lesotho is sometimes described as a "hostage state." It is a member of the Commonwealth and of the African Union, and it is a signatory of the Lomé Convention. Before its independence on Oct. 4, 1966, it was one of the three British High Commission Territories, the other two being Bechuanaland (now Botswana) and Swaziland.

Lesotho is the name of the country, and Basotho (Sotho) is the name of its people; a single individual is referred to as a Mosotho. Sesotho is the language. The capital of Lesotho is Maseru.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Two-thirds of the country consists of mountains. The highest peak, Mount Ntlenyana, is 11,424 feet (3,482 metres) above sea level. The Drakensberg range forms the eastern boundary with KwaZulu/Natal. The Maloti spurs, running north and south, join the main range in the north, where they form a plateau between 9,000 and 10,500 feet in elevation. This plateau is the source of South Africa's two largest rivers—the eastward-flowing Tugela and the westward-flowing Orange—as well as tributaries of the Caledon (Mohokare). The foothills, with elevations averaging between 6,000 and 7,000 feet, descend in undulating slopes to the west, where the lowlands bordering Free State average some 5,000 to 6,000 feet in elevation.

The plateau

The mountain soils are of basaltic origin and are shallow but rich. The soils of the lowlands derive mainly from the underlying sandstone. Extensive erosion has severely damaged soils throughout the country.

Climate. Precipitation, brought by the prevailing winds, occurs mostly between October and April and is variable; the annual average is about 28 inches (710 millimetres), with amounts decreasing from east to west. Although droughts are rare, their periodic occurrence is devastating to the country. Temperatures in the lowlands vary from 90° F (32° C) in the summer to 20° F (−7° C) in the winter. In the highlands the temperature range is much wider, and readings below 0° F (−18° C) are not unusual. Frost occurs widely in the winter, when the Maloti are usually snowcapped. Hail is a frequent hazard during the summer.

Plant and animal life. Overgrazing, overutilization, and soil erosion have severely depleted and altered the grasslands, the reedbeds, and the woody bush on the protected slopes and led to the invasion of unpalatable vegetation over wide areas. Indigenous trees include the Cape willow (*Salix capensis*) and wild olive, and the wild willow and white poplar have been introduced. Afforestation schemes (the Woodlot Project) have been attempted. One of the eight indigenous species of aloe is the endemic spiral aloe (*Aloe polyphylla*).

Large mammals have been eradicated, but smaller antelope and hares can be found, and the hyrax, or dassie, is common. Lesotho is the last stronghold in southern Africa of the magnificent bearded vulture, or lammergeyer (*Gypaetus barbatus*). Rivers contain yellowfish, the rare Maloti minnow (*Oreodaimon quathlambae*) is of interest, and trout have been introduced.

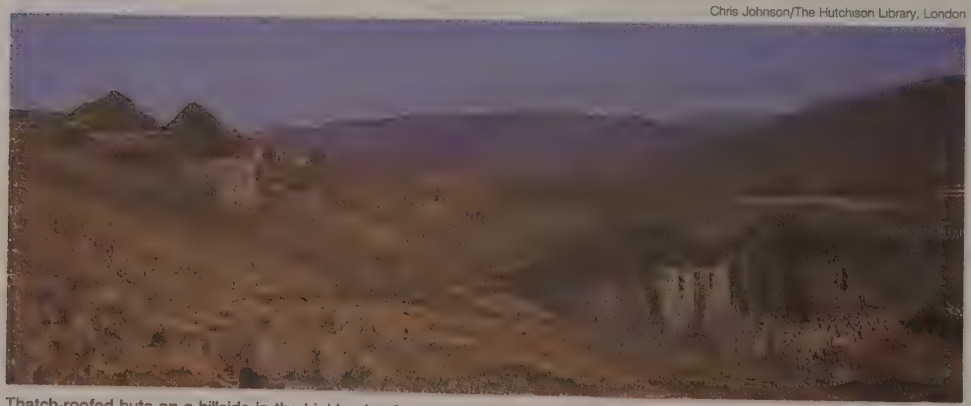
Settlement patterns. Although not permanently inhabited, the mountain grasslands on the slopes of the high plateau as well as in the valleys provide summer grazing for sheep and cattle, which are tended by herders in isolated cattle posts. Some of the deep valleys, like the Senqunyane, produce crops of wheat, peas, and beans. One-half of the population lives in the lowlands, which form a narrow corridor, averaging only 25 miles (40 kilometres) in width, along the Caledon River.

The Basotho combine modern and traditional ways to an unusual degree; this provides continuity in a society that is disrupted by a system of migratory labour. Although undermined by political developments since independence, traditional authority is still exercised through a system of chieftaincy extending from the king through the chiefs to the village level. Their authority rests largely on their responsibility for the working and distribution of the land, although in certain areas this authority has been reduced by the Land Act of 1979.

Since independence there has been considerable movement toward the capital city, Maseru, the population of which doubled between 1966 and 1976. It now consists of a developed modern city centre surrounded by suburbs for the large bureaucracy and for foreign aid and development personnel, with informal settlements on the periphery. The Maseru Municipal Council was constituted in 1988.

Social cohesion is strengthened by the persistence of clan and family loyalties. Families and clans still cluster together as units in the numerous small rural villages. There are no large towns in the country. The villages range in size from one large family to four or five extended families, with an average of 30 to 50 immediate families. The villages are often picturesque, offering fine views of the rocky highlands; on the plains they are often surrounded by aloes

Chris Johnson/The Hutchison Library, London



Thatch-roofed huts on a hillside in the highlands of central Lesotho, near Semonkong

and trees, and the walls and doors of many houses are covered with patterned designs. The villages themselves consist of clusters of circular or rectangular huts solidly built of turf, Kimberley brick (unburned clay), or dressed stone, mostly with thatched roofs; more recently, however, corrugated iron has become a popular roofing material.

The average household usually has two or three huts, the larger one being used as a living and dining area and as the parents' bedroom; the smaller ones are used for kitchen and storage purposes and as sleeping quarters for the children. The hut of the chief, or headman, is usually in the centre of the village, flanked by that of the principal wife and ringed by those of the junior wives. The *lekhotla* (open court) is in front of the chief's hut; beside it are the kraals (enclosures) for the cattle and stables for horses. Village life centres largely on the fields, the chief's court, the kraals, the school, the church, and the initiation lodge. Circumcision forms an integral part of the ritualized initiation ceremonies that train boys to take their place as full members of the family, clan, and nation—the three centres of social cohesion. Young boys still spend a large part of their lives as herdsmen, while women and young girls do much of the hard work in the fields, a practical necessity because the able men are usually absent, although they contrive to return home briefly for the plowing or the harvest.

Initiation ceremonies

Since independence the building of hard-top roads in the lowlands and the opening of good-quality gravel roads into the highlands, together with the availability of four-wheel-drive road vehicles and domestic flights, have done much to break down the isolation of the more remote villages. However, the small, sturdy Basotho pony is still widely used in the rural areas and is often the only means of rural transport. Because of the sharp variations in climate, both men and women invariably wear blankets, which they use as cloaks; a great deal of care is taken in choosing a blanket, which, especially for the men, is usually multicoloured. Men and women also wear the typical Sotho hat, which is woven from reed into conical shapes, with an unusual topknot.

The people. *Ethnic groups.* The Basotho are members of the Southern Sotho linguistic group, originally united by a common loyalty to the royal house of Moshoeshoe (Mshweshwe), of the Moketeli branch of the Kwena lineage. Internally, cleavages between different chiefdoms (and within the royal lineage itself) have had political significance, but externally the sense of Basotho nationhood and cultural unity remains strong.

There are a few thousand nationals of Asian or Coloured (mixed race) descent, but the European community is dominated by expatriate teachers, missionaries, aid workers, technicians, and development advisers.

Emigration

Because the country is severely overpopulated, both temporary and permanent emigration has taken place. No exact figures exist for the number of Basotho working or living in South Africa. The great majority of the migrant workers are men under the age of 40.

Religion. The majority of Basotho profess Christianity and are members of the Roman Catholic church, the Lesotho Evangelical church, or the Anglican church. Independent churches are also present, together with Zionist sects. Traditional beliefs are still held but are of declining significance.

Demographic trends. The population density of Lesotho is high for an African state, despite the country's mountainous terrain, but the population is growing at a much slower rate than that of most other states. Most of the population live in the western lowlands, which have a population density four times higher than the country as a whole.

Labour migration dominates the life of the Basotho, with more than one-half of the adult male labour force working in South Africa.

The economy. Lesotho is a poor country with few natural resources; these are certainly insufficient for even the present population. Its economy could not be sustained at all without such benefits as it derives from South Africa, with which Lesotho forms part of a customs union and shares an integrated communications system. Lesotho also depends wholly on South Africa for the export of its surplus working population.

Economic links to South Africa

The government is the largest employer of labour in the country, with a large share of its annual budget being made up of payments to its public employees. Three forms of direct taxation are levied: a basic tax on all adult males; a graded tax on all income earners; and an income tax. There is also a general sales tax.

Agriculture. Only 10 percent of Lesotho is arable, but agriculture forms the primary occupation and is the single largest contributor to the gross domestic product (GDP). However, its contribution has declined to less than a quarter of total GDP. The most important agricultural products are corn (maize), sorghum, wheat, beans, and peas. Cattle products have been exported, but cattle have more social than commercial value. Wool and mohair are produced and exported. Agricultural development projects are funded by a wide range of agencies, including the World Bank. None, however, have been able to reverse the steady decline in agricultural production since the mid-1960s.

Industry. Geologic surveys have so far shown little promise of mineral wealth, although kimberlite pipes in the highlands do produce diamonds. The capital-intensive mine at Lestseng-la Terai, opened by DeBeers in 1977, was closed in 1982.

Development of industry and manufacturing has been the responsibility of the Lesotho National Development Corporation and the Basotho Enterprises Development Corporation. Three industrial estates operate small projects; among their products are candles, ceramics, furniture, and jewelry, and they include weaving, canning, and diamond-polishing concerns. Urban development has stimulated the construction, catering, and service sectors. Although the manufacturing and construction sectors have grown markedly, they provide less than 10 percent of all jobs. Together they contribute some 25 percent to the GDP, while manufacturing contributes half of all export earnings.

The Highlands Water Scheme

The Highlands Water Scheme is a significant project for the future of the region as a whole and Lesotho in particular. Through this project the headwaters of the Orange River in the deep valleys of the Lesotho highlands are to be dammed, and the immense volume of water this will accumulate will be drained off and piped to the industrial regions of South Africa. Included in the plan is a hydroelectric scheme for Lesotho. It is a long-term project that will take at least 30 years to complete in four stages. The project is run by a Joint Standing Commission and has attracted backing from the World Bank, the European Economic Community, and a number of other development agencies.

Tourism. The mountain scenery of Lesotho has potential for tourism. Roads and pony trails have been developed, trout streams stocked, and hotels built in Maseru and other places of interest to exploit this market. Tourism in Lesotho will receive a considerable boost with the development of the Highlands Water Scheme.

Finance. Lesotho has been a member since 1986 of the Tripartite Monetary Area, comprising Lesotho, Swaziland, South Africa, and Namibia. Its currency, the loti (plural, maloti), is fixed to the South African rand. Lesotho is a member of the South African Customs Union, and payments from the union make up a substantial proportion of government revenue. An economic program negotiated with the World Bank led to the granting of a substantial IMF loan under its structural adjustment facility in 1988.

Trade. Lesotho has a large and growing trade deficit. Most trade is with South Africa. The large deficit is offset by the remittances of Lesotho's migrant workers, by external aid, and by receipts from the customs union in agreement with South Africa, Swaziland, and Botswana.

Transportation. A main road runs along the western and southern boundary, and a mountain road from Maseru reaches into the interior. These two main arteries are served by short-distance feeder roads. Villages in the mountains are served by bridle paths. A one-mile railway line links the capital to the South African transport network.

Considerable use is made of light aircraft for passengers

and for transporting mail and freight to the interior. An international airport south of Maseru became operational in 1986.

Administration and social conditions. *Government.* In July 1991 a new constitution was approved by a constituent assembly. Promulgated after the March 1993 general election, it calls for a bicameral parliamentary system with the king as hereditary head of state. The National Assembly has 65 directly elected members; the Senate has 33 members (22 principal chiefs and 11 appointees).

The old constitution was suspended following the disputed 1970 election, in which the opposition claimed victory over the ruling Basotho National Party (BNP). The legislature was replaced by an appointed 93-member interim assembly that included the 33 members of the Senate. In 1973 a new National Assembly, consisting of the 22 principal chiefs and 71 members from the BNP and other parties, was appointed. It was dissolved in 1986, and elections were not held again until 1993.

Basotho National Party

Justice. The legal system is based on Roman-Dutch law, with elements of British and customary law playing a role. There are local and central courts, judicial commissioners' courts, subordinate courts, and a Court of Appeal, with the High Court as a superior court of record.

Education. Although Lesotho has one of the highest literacy rates in Africa, the educational structure does not have great depth, and only a small percentage of students have the resources and skills to reach the higher levels. Education remains largely the responsibility of the churches, under the supervision of the Ministry of Education. Considerable funding and expertise have contributed to the development of a more relevant and adequate system, but some unbalanced features remain, with the National University of Lesotho, for example, absorbing about one-fifth of recurrent expenditure. There are also agricultural and educational training centres.

Health and welfare. The country's healthy climate contributes to the comparatively low rate of sickness. The main incidence of ill health is the result of food-deficiency diseases, venereal disease, chronic rheumatism, infections of the respiratory tract, and dyspepsia. There are several hospitals, about half of which are operated by the government, and a number of clinics for maternity, child health, and venereal diseases, as well as a few health centres and dispensaries. In recent years the demands of the IMF structural adjustment program have threatened to reduce the capacity of some of the country's leading health programs and institutions.

Cultural life. The contradictions created by Lesotho's political independence and economic dependence are reflected in the cultural life of the country. Despite increasing urbanization and the growth of modern institutions and bureaucracy (including the experience of many Basotho in industry), the overall objective is to build the rural homestead and perpetuate traditional institutions. The paramount chief and the system of chieftaincy remain a focus of loyalty, although they have been increasingly discredited in recent years. Institutions such as the initiation schools, which perpetuate traditional values, are still significant although they are changing in structure and declining in importance.

Changes since independence

Although rural conservatism remains a feature, the pace of change in Lesotho has been marked since independence. With the process of democratization advancing in neighbouring South Africa and the development of the Highlands Water Scheme, these changes will intensify.

Nonetheless the historical traditions and legacy of Moshoeshoe I, founder of the nation, remain strong, and there is national pride in Lesotho's history of resistance, the role of the Basotho in building modern southern Africa, and the achievements of its writers (Thomas Mofolo) and composers (Joshua Pulumo Mohapeloa). The newspaper *Leselinyana La Lesotho* has been published for more than a century, and the printing presses of the mission stations, such as the Morija Press, made a substantial contribution to the educational literature of southern Africa.

Village life is dominated by basic agricultural tasks, with heavy responsibilities falling on women. Craftwork is still practiced in the villages and includes pottery and grass-weaving (which is most obviously seen in the traditional Basotho hat), and the walls of houses are often attractively decorated.

Herders still play the traditional *letsiba,* and dances such as the "gum-boot dance" and the *lefela* demonstrate the influence of the experience of labour migrancy on traditional forms of cultural expression.

The radio has done much to improve communication in Lesotho, where broadcasts from South African stations can be received; the British Broadcasting Corporation (BBC) has built a regional transmitter. Broadcasts in Sesotho are received from the Lesotho National Broadcasting Service.

For statistical data on the land and people of Lesotho, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR. (Co.L./J.J.Gu.)

HISTORY

The territory now known as Lesotho has been occupied from the dawn of human history and during the past few thousand years by Late Stone Age San hunter-gatherers. From about the 16th century, African farmers—the ancestors of the present population—moved across the grasslands of southern Africa and settled in the fertile valleys of the Caledon River, where they came to dominate the San hunters of the region. These stock-keeping agriculturalists belonged to the large Sotho group and were divided into numerous clans that formed the nucleus of chiefdoms whose members occupied villages and the associated arable and grazing land.

The Basotho kingdom (1824–69). The violent upheavals of the early 19th century among the chiefdoms of southern Africa became particularly intense in Lesotho in the 1820s. During this violence, known as the *lifaqane* ("hammering"; also spelled *difaqane*), the members of many chiefdoms were annihilated, dispersed, or incorporated into reorganized, larger chiefdoms in strategically advantageous areas; these were under new leaders who had the ability to offer greater protection.

One such man was Moshoeshoe (Mshweshwe; b. *c.* 1786) of the Moketeli, a minor lineage of the Kwena (Bakwena). In 1824 he occupied the mountain Thaba Bosiu ("Mountain of the Night"), the defensive centre from which he incorporated many other individuals, lineages, and chiefdoms into what became the kingdom of the Basotho. Moshoeshoe was a man of remarkable political and diplomatic skill. By cooperating with other chiefdoms and extending the influence of his own lineage, he was able to create a sense of Basotho identity and unity with which to confront the external forces that soon posed a threat to their autonomy and independence. He was also aware of the need to acquire the skills of the increasing number of farmers, settlers, hunters, and adventurers moving across his borders from the south if he was to counter the challenges they posed. He therefore welcomed the missionaries from the Paris Evangelical Missionary Society when they arrived at Thaba Bosiu in 1833 as a source of information about the rest of the world, relating its ideas and material achievements. He placed them in strategically important parts of the kingdom, where they gave the Basotho their first experience of Christianity, literacy, and commodity production.

Moshoeshoe

External influence was changed to external threat in the late 1830s when large numbers of Boer trekkers from the Cape Colony settled on the western margins of the kingdom and began to contest the right of the Basotho to their land. The next 30 years were characterized by conflict between the Basotho and the Boers, which at times broke out into overt warfare and led ultimately to the loss of most Basotho land to the west of the Caledon River. The third party in these wars was the British, to whom Moshoeshoe appealed for intervention. The British governors of the Cape redrew the boundaries between the Boers and the Basotho on a number of occasions but failed to resolve the conflict.

After devastating wars in the late 1860s Moshoeshoe, fearing the dispersal and possible extinction of his people, appealed to the British for assistance. Sir Philip Wodehouse, governor and high commissioner of the Cape Colony, concerned for the stability of the region and

British interests in southern Africa, annexed the kingdom to the British crown. The British government endorsed his action, granting the new colony, Basutoland, to the Cape Colony in 1871.

Moshoeshoe died in 1870 and was buried on Thaba Bosiu. The "heroic age" of the Basotho had come to an end. The independent African mountain kingdom, forged against such odds in the earlier part of the 19th century, had lost much of its most productive land to the Boers and its political autonomy to the British. Nonetheless, the Basotho had retained some of their land and also their social and cultural independence, together with a fierce sense of injustice based on the memory of their resistance and what they had lost to their neighbours—a memory which they carried with them into the 20th century.

Basutoland (1871–1966). Attempts by the Cape Colony administration to co-opt the chieftainship and disarm the Basotho led to the Gun War of 1880–81. The Cape Colony relinquished Basutoland to British rule in 1884, when it became one of the British High Commission Territories in southern Africa.

Meanwhile, developments had been occurring in southern Africa that were to determine the course of Basutoland's future. The first mineral discoveries had been made that were to lay the foundation for the creation of the Union of South Africa (1910). In pursuit of racial supremacy, cheap labour, and the end to competition from independent African agricultural producers, South Africa deprived the African population of its social and political rights and most of its land. The small British-ruled enclave of Basutoland was profoundly affected by these policies. Basotho farmers took advantage of the markets for foodstuffs in the growing South African mining centres, and during the latter part of the 19th century they utilized new farming techniques to produce substantial surpluses of grain, which they sold on the South African markets. With the discovery of minerals, Basotho workers traveled to the mines to sell their labour for cash and for firearms.

Effects of the Union of South Africa

Increasing dependency on labour migrancy to South Africa has been the dominant feature of Lesotho's history in the 20th century, made necessary by taxation, increasing population behind a closed border, depletion of the soil, and the need for resources to supplement agricultural production and to acquire the means to purchase livestock for bridewealth. Basotho workers became an important element of the South African mining industry and Basutoland the classic example of the southern African labour reserve, its people dependent on work in South Africa for their survival.

By setting up a system of dual rule the British left considerable powers in the hands of the chiefs. They were headed by a paramount chief: Letsie (1870–91), Lerotholi (1891–1905), Letsie II (1905–13), Griffith (1913–39), and Seeiso (1939–40)—all of whom were descendants of Moshoeshoe—and subsequently the regent 'Mantsebo (1940–60). Under the paramount chiefs, authority was delegated through ranked regional chiefs drawn from the royal lineage and the most important chiefdoms. A system of customary law was adopted, with the land held in trust by the paramount chief for the people, and crucial aspects of local government were in the hands of the chiefs. The colonial government was headed by a resident commissioner and advised by the Basutoland National Council, which was led by the paramount chief and dominated by his nominated members.

The British administration was concerned primarily with balancing Basutoland's budget, and this was facilitated by ensuring that a substantial proportion of the population worked for wages in South Africa. The often domineering local chiefs could do little to halt the increasing social and economic deprivation within Basutoland. Education was left to the missionary societies, and there was little development of economic infrastructure or social services. The effects of Basutoland's external linkages with the developed world, and its internal underdevelopment, were seen clearly between 1929 and 1933, when the Great Depression coincided with a massive drought in Basutoland, driving so many people into South Africa that the population hardly increased for a decade.

Opposition to colonialism

Opposition to the colonial system grew. The missionary societies had created a vocal and informed class of teachers and writers. The Basutoland Progressive Association (founded in 1907) worked to reduce chiefly power and for representation on the Basutoland National Council; Lekhotla la Bafo (1919) was a more radical, populist organization. Neither organization, however, was able to effectively counter the power of the colonial adminstration and its traditionalist allies. An important factor that restricted all opposition movements and united all Basotho was their opposition to South Africa and their fear that the British might implement their intention to cede Basutoland to South Africa.

Widespread criticism of the system of administration led to attempts from the 1930s to reduce the number and prerogatives of the chiefs, but after World War II attempted reform was outstripped by the development of nationalist parties pressing for independence. Three major political parties emerged at this time: the Basutoland Congress Party (BCP) in 1952, under Ntsu Mokhehle; the more conservative Basutoland National Party (BNP) in 1958, under Chief Leabua Jonathan, which was associated with chiefly power and the Roman Catholic church; and the Marema-Tlou Freedom Party (1963), which was identified more with the defense of the powers of the country's principal chiefs, in particular the paramount chief, Bereng.

In the 1965 elections the BNP, with strong support from the Roman Catholic church and South African and white trading interests, succeeded in gaining a narrow majority over the BCP. On Oct. 4, 1966, Basutoland received its independence from Britain as the Kingdom of Lesotho, with Moshoeshoe II, named after the nation's founder, as the paramount chief.

The Kingdom of Lesotho. Lesotho faced tremendous problems. The country was poor and overpopulated, and the land was overworked. Agricultural production was low and had to be supplemented by massive labour migrancy to, and imports from, South Africa. Apart from land and labour, the only potential local resources for development were water and the diamonds in the mountains. The government was dependent on grants from Britain, the payments it received from South Africa as a member of the South African Customs Union, and the earnings brought into Lesotho from South Africa by migrant workers.

Chief Jonathan's regime

In January 1970 the first postindependence general elections were held, and the opposition BCP gained a majority of seats. But the results were never released, and Chief Jonathan suspended the constitution, arrested leading members of the opposition, and sent the king into temporary exile. Resistance to these moves was put down with considerable violence. After a short delay Britain recognized the legitimacy of the new regime.

The ruling party used legislation and violence—and the distribution of state patronage—to silence and control its opponents. In 1974 the BCP attempted to overthrow the regime, but this coup was put down, and Mokhehle of the BCP went into exile.

In an attempt to gain further acceptance and support within the international community, Chief Jonathan and his government began to adopt an increasingly hostile attitude toward its powerful neighbour, South Africa. The idea of supporting a small, embattled, sovereign African country in its struggle against its racist neighbour was attractive to a large number of donor countries, and during the 1970s Lesotho received an increasing amount of foreign aid. The overall assessment of these projects is that—while they attracted a significant amount of funding to Lesotho and increased the pace of modernization, urban development, economic infrastructure, education, communications, and the creation of a privileged bureaucracy—they failed to alleviate the long-standing problems of poverty and dependence. Thus, although mine wages and payments from the South African Customs Union increased in the 1970s, Lesotho was unable to use the increased revenues productively and remained dependent on South Africa.

The Lesotho government's hostility to the South African government and its acceptance of South African refugees also began to have serious consequences. As part of its

strategy of destabilizing its African neighbours, South Africa gave support to the armed wing of the BCP, the Lesotho Liberation Army, and in December 1982 the South African Defence Force attacked houses in Maseru that it alleged were African National Congress guerrilla bases, killing more than 40 people, many of whom were citizens of Lesotho. Relations grew increasingly tense as the South African government demanded the expulsion of South African refugees and began to interfere with the free flow of people and goods moving across the border.

Meaningful political activity remained confined to the upper social echelons, and differences began to appear among leading figures within the government; one faction advocated a policy more amenable to South African demands. Matters came to a head in January 1986 when the South African authorities placed severe restrictions on the movement of goods and people across the border, effectively closing it. Confronted with the consequences of this blockade, the pro-South African faction, led by Major General Justin Lekhanya, deposed Chief Jonathan (who died the next year) and took power in the name of the king and national reconciliation, establishing military rule with the king as head of state.

The Military Council

The Military Council banned open political activity and deported a number of South African refugees; South Africa lifted the blockade. In October 1986 Lesotho and South Africa signed the Lesotho Highlands Water Treaty, and in 1987 a South African trade mission was established in Lesotho.

Internally the Military Council did little to resolve political tensions, and it became associated with a number of brutal deaths and assassinations. Lesotho's economic impasse continued as recession and the general crisis deepened in South Africa and the gold mining industry reduced production. Conditions imposed by IMF loans also caused economic hardships. (J.J.Gu.)

In February 1990 conflict arose within the ruling class, and King Moshoeshoe went into exile. His eldest son, Mohato, was sworn in as Letsie III. In April 1991 Lekhanya was forced to resign, and his replacement lifted the ban on political activity. The political and economic crises continued, however, and in the general elections of March 1993, the BCP came to power under the leadership of Ntsu Mokhehle. Letsie's attempt to dismiss Mokhehle's government in August 1994 proved unsuccessful, and Moshoeshoe was reinstated as king in January 1995. Less than a year later, Moshoeshoe died, and Letsie reassumed the throne.

In 1997 the BCP dismissed Mokhehle as leader, and he eventually formed his own party, the Lesotho Congress of Democrats (LCD), which overwhelmingly won the general elections of May 1998. Upon Mokhehle's resignation, Pakalitha Mosisili became prime minister. The election was declared free and fair by many international observers, but the government faced an insurrection and asked the South African Development Community (SADC) to send troops to quell it. Eventually, the SADC forces restored order, and South Africa imposed an agreement that called for new elections by mid-2000. Elections were postponed until 2002. Meanwhile, the LCD government that took power in May 1998 under Mosisili signed an agreement with representatives from the SADC that put into effect an Interim Political Authority (IPA) with opposition representation. (Ed.)

For later developments in the history of Lesotho, see the BRITANNICA BOOK OF THE YEAR.

Malawi

A landlocked southeastern African country of spectacular highlands and extensive lakes, the Republic of Malawi occupies a narrow, curving strip of land along the East African Rift Valley. Stretching about 520 miles (840 kilometres) from north to south, it has a width varying from 5 to 100 miles and is bordered by Tanzania to the north, Mozambique to the east and south, and Zambia to the west. Its total area of 45,747 square miles (118,484 square kilometres) includes 9,347 square miles of lake surface dominated by the 8,900 square miles of Lake Nyasa (known in Malawi as Lake Malawi). Lilongwe is the seat of the legislature, and Blantyre is the seat of the judiciary and the executive.

Most of Malawi's population engages in cash-crop and subsistence agriculture. The country's exports consist of the produce of both small landholdings and large tea and tobacco estates. Malawi has successfully attracted foreign capital investment, has made great strides in the exploitation of its natural resources, and is one of the few African countries to regularly produce food surpluses.

PHYSICAL AND HUMAN GEOGRAPHY

The four basic regions

The land. *Relief.* While Malawi's landscape is highly varied, four basic regions can be identified: the East African (or Great) Rift Valley, the central plateaus, the highlands, and the isolated mountains. The East African Rift Valley—by far the dominant feature of the country—is a gigantic troughlike depression running through the country from north to south and containing Lake Malawi (north and central) and the Shire River valley (south). The lake's littoral, situated along the western and southern shores and ranging from 5 to 15 miles in width, covers about 8 percent of the total land area and is dotted with swamps and lagoons. The Shire valley stretches some 250 miles from the southern end of Lake Malawi at Mangochi to Nsanje at the Mozambique border and contains Lake Malombe at its northern end. The plateaus of central Malawi rise to an elevation of 2,500 to 4,500 feet (760 to 1,370 metres) and lie just west of the Lake Malawi littoral; the plateaus cover about three-fourths of the total land area. The highland areas are mainly isolated tracts that rise as much as 8,000 feet above sea level. They comprise the Nyika, Viphya, and Dowa highlands and Dedza-Kirk Mountain Range in the north and west and the Shire Highlands in the south. The isolated massifs of Mulanje (9,849 feet) and Zomba (6,841 feet) represent the fourth physical region. Surmounting the Shire Highlands, they descend rapidly in the east to the Lake Chilwa–Phalombe plain.

Drainage and soils. The major drainage system is that of Lake Malawi, which covers some 11,430 square miles and extends beyond the Malawi border. It is fed by the North and South Rukuru, Dwangwa, Lilongwe, and Bua rivers. The Shire River, the lake's only outlet, flows through adjacent Lake Malombe and receives several tributaries before joining the Zambezi River in Mozambique. A second drainage system is that of Lake Chilwa, the rivers of which flow from the Lake Chilwa–Phalombe plain and the adjacent highlands.

Soils, composed primarily of red earths, with brown soils and yellow gritty clays on the plateaus, are distributed in a complex pattern. Alluvial soils occur on the lakeshores and in the Shire valley, while other soil types include hydromorphic (excessively moist) soils, black clays, and sandy dunes on the lakeshore.

Climate. There are two main seasons—the dry season from May to October and the wet season from November to April. Temperatures vary seasonally as well, and they tend to decrease on average with increasing elevation. Nsanje (Port Herald), in the Shire River valley, has a mean July temperature of 69° F (21° C) and an October mean of 84° F (29° C), while Dedza, which lies at an altitude of more than 5,000 feet, has a July mean of 57° F (14° C) and an October mean of 69° F (21° C). On the Nyika Plateau and on the upper levels of the Mulanje Massif, frosts can occur in July. Annual rainfall is highest over parts of the northern highlands and on the Sapitwa peak of Mulanje Mountain, where it is about 90 inches (2,300 millimetres); it is lowest in the lower Shire valley, where it ranges from 25 to 35 inches (650 to 900 millimetres).

Diversity of vegetation

Plant and animal life. The natural vegetation pattern reflects the country's diversity in elevation, soils, and climate. Savanna (grassy parkland) occurs in the dry lowland areas. Open woodland with bark cloth trees, or *miombo* (leguminous trees unsuitable for timber), is widespread on the infertile plateaus and escarpments. Woodland, with species of acacia tree, covers isolated, more fertile plateau sites and river margins; grass-covered, broad depressions, called *madambo* (singular: *dambo*), dot the plateaus; grassland and evergreen forest are found in conjunction on the highlands and on the Mulanje and Zomba massifs.

Malaŵi's natural vegetation, however, has been altered significantly by human activities. Swamp vegetation has given way to agricultural species as swamps have been drained and cultivated. Much of the original woodland has been cleared, and, at the same time, forests of softwoods have been planted in the highland areas. High population density and intensive cultivation of the Shire Highlands have also hindered natural succession there, while wells have been sunk and rivers dammed to irrigate the dry grasslands for agriculture.

Game animals abound only in the game reserves, where antelope, buffalo, elephants, leopards, lions, rhinoceroses, and zebras occur; hippopotamuses live in Lake Malaŵi. The lakes and rivers contain more than 200 species and 13 families of fish. The most common and commercially significant fish include the endemic tilapia, or *chambo* (nest-building freshwater fish); catfish, or *mlamba;* and minnows, or *matemba.*

Settlement patterns. Malaŵi is the most densely populated country in southern Africa, but ironically it is also one of the least urbanized, with 9 out of 10 people living in rural locations. A rural village—called a *mudzi*—is usually small. Organized around the extended family, it is limited by the amount of water and arable land available in the vicinity. On the plateaus, which support the bulk of the population, the most common village sites are at the margins of *madambo,* which are usually contiguous with streams or rivers and are characterized by woodland, grassland, and fertile alluvial soils. In highland areas, scattered villages are located near perennial mountain streams and pockets of arable land. The larger settlements of the Lake Malaŵi littoral originated in the 19th century as collection points for slaves and later developed as lakeside ports. Improvements in communication and the sinking of wells in semiarid areas have permitted the establishment of new settlements in previously uninhabited areas. Architecture is also changing; the traditional round, mud-walled, grass-roofed hut is giving way to rectangular brick buildings with corrugated iron roofs.

Urban development began in the colonial era with the arrival of missionaries, traders, and administrators and was further stimulated by the construction of the railway. The only true urban centres are Blantyre-Limbe, Zomba, Mzuzu, and Lilongwe. Although some district centres and missionary stations have an urban appearance, they are closely associated with the rural settlements surrounding them. Blantyre, Malaŵi's industrial and commercial centre, is situated in a depression on the Shire Highlands at an altitude of about 3,400 feet. Zomba, seat of the University of Malaŵi, lies at the foot of Zomba Mountain and is purely of administrative origin. Farther north is Lilongwe, Malaŵi's new capital, which is developing agricultural industries.

The people. Nine major ethnic groups are historically associated with modern Malaŵi—the Chewa, Nyanja, Lomwe, Yao, Tumbuka, Sena, Tonga, Ngoni, and Ngonde (Nkonde). All the African languages spoken belong to the Bantu language family. Chichewa is the national language and English the official language, although English was understood by less than one-fifth of the population at independence. Chichewa is spoken by about two-thirds of the population. Other important languages are Chilomwe, Chiyao, and Chitumbuka.

Predominant languages

Some two-thirds of the population are Christian, of which more than half are members of various Protestant denominations and the remainder Roman Catholic. Muslims constitute almost one-fifth of the population, and traditional beliefs are adhered to by nearly everyone else.

The population is growing at a rate well above average for sub-Saharan Africa. The birth rate is one of the highest on the continent, but the death rate is also high, and life expectancy—at 47 years—is significantly below average for a southern African country. With nearly one-half the population younger than age 15, high birth and population-growth rates should continue in the 21st century. By the early 1990s the problem of high population growth was compounded further by the then decade-long influx of refugees from Mozambique—estimated to number about one million—fleeing the civil war in that country.

The economy. The backbone of the Malaŵi economy is agriculture, which regularly accounts for one-third of the gross domestic product and 90 percent of export earnings and which employs more than 80 percent of the working population. Since the mid-1960s, however, the sector has become increasingly concentrated on three cash crops—tobacco, tea, and sugar—and increasingly dependent on the market demand for these commodities. The small industrial sector is geared largely to processing agricultural products, with some limited manufacturing of import substitutes.

The government has sought to strengthen the agricultural sector by encouraging integrated land use, higher crop yields, and irrigation schemes. In pursuit of these goals, several large-scale integrated rural development programs, covering one-fifth of the country's land area, have been put into operation. These projects include extension services; credit and marketing facilities; physical infrastructures such as roads, buildings, and water supplies; health centres; afforestation units; and crop storage and protection facilities. Outside the main program areas, advisory services and educational programs are available, and the Malaŵi Young Pioneers, a national youth movement, trains more than 2,000 young men and women yearly in techniques of rural development.

Promoting higher crop yields

Both higher incomes in the rural areas and continued public expenditure are factors that government planners hope will increase the purchasing power of the public as a whole and thus provide a stimulus for further industrial development. The government continues to promote the establishment of import-substitute industries, in hopes of reducing reliance on expensive imported goods, strengthening the balance-of-payments situation, and, at the same time, increasing employment opportunities.

Resources. Most of Malaŵi's mineral deposits are neither extensive enough for commercial exploitation nor easily accessible. Some small-scale mining of coal takes place at Livingstonia and Rumphi in the north, and quarrying of limestone for cement production is also important. Exploration and assessment studies continue on other minerals such as apatite, located south of Lake Chilwa; bauxite, on the Mulanje Massif; kyanite, on the Dedza-Kirk Range; vermiculite, south of Lake Malaŵi near Ntcheu; and rare-earth minerals, at Mount Kangankunde northwest of Zomba. Deposits of asbestos, uranium, and graphite are known to exist as well.

More than half of Malaŵi's total land area is potentially arable, though only about one-fourth of it is cultivated regularly. Forests and woodlands cover nearly half of the country, and almost 4,000 square miles are in state-controlled forest reserves.

The lakes and rivers of Malaŵi are estimated to provide more than 60 percent of the country's animal protein intake. Lake Malaŵi, in particular, is a rich source of fish within easy access for most of the country's population.

Malaŵi's water resources are plentiful, although some rural areas are inadequately supplied. Treated water for the major cities of Blantyre and Lilongwe is supplied by the Walker's Ferry Scheme and the Kamuzu Dam, respectively. Most of the rivers are seasonal, but a few large ones, particularly the Shire River along its middle course, have a considerable irrigation and electricity-generating potential. The total hydroelectric potential of the country is estimated to be about 1,200 megawatts, of which more than 500 megawatts can be generated on the Shire River alone. Present power demands, which represent only about 10 percent of potential capacity, are met by the Nkula Falls (two plants) and Tedzani Falls hydroelectric schemes and by diesel plants.

Agriculture, fishing, and forestry. The most important agricultural export products are tobacco, tea, sugar, and peanuts (groundnuts). Tea is grown on plantations on the Shire Highlands by the largest proportion of the country's salaried labour force. Tobacco, by far the most important export, is raised largely on the central plateau on large estates. Corn (maize) is the principal food crop and is typically grown with beans, peas, and peanuts throughout the country by virtually all smallholders. Other important crops are cotton, cassava, coffee, and rice. Although the

Importance of tobacco, tea, and sugar

River fish being dried in the open air in the lower Shire River valley, southern Malaŵi. A baobab tree stands in the background to the right.
Andrew C. Millington

major share of commercial crop production and nearly one-fifth of all cultivated acreage is on large estates, most farms are small, averaging less than 3 acres (1.2 hectares). Smallholder cash crops are purchased and marketed by the Agricultural Development and Marketing Corporation; a few cooperative societies purchase and market produce.

Lake Malaŵi is the major source of Malaŵi's fishing industry, but Lakes Chilwa and Malombe and the Shire River also contribute significantly to the annual catch. The industry supplies mainly a local market, but some fish are exported to neighbouring countries.

Since the early 1970s the government has sponsored the development of several large timber and pulpwood plantations aimed at making the country self-sufficient in construction grades of timber. Pine and eucalyptus have also been planted extensively in the northern Viphya Mountains to supply a large pulp and paper project in the region. Sawn poles, posts, and manufactured wooden items are produced largely for the domestic market, although some forest products are exported.

Industry. Development of the country's industrial base was accorded high priority at independence, and Malaŵi now satisfies much of its domestic need for products such as cotton textiles, canned foodstuffs, beer, edible oils, soaps, sugar, radios, hoes, and shoes, all of which previously had to be imported. The main demand for electric power is in the industrial areas of the south near Blantyre, where electricity consumption has steadily multiplied; the industrial area of Lilongwe; the vast sugar estates of Sucoma and Dwangwa; and the pulpwood scheme of Viphya.

Bauxite production

Exploitation of bauxite, Malaŵi's most economically important mineral reserve, will depend on an increased hydroelectric capacity to meet the demand of bauxite smelting for abundant cheap electric power. Only such an energy supply could offset the heavy costs of transporting the ore from its remote location to be processed into alumina and then of transporting the alumina to the coast for export.

Trade and finance. About two-thirds of Malaŵi's foreign-exchange earnings are derived from exports of tobacco, of which Malaŵi is the second largest producer in Africa (after Zimbabwe). The main purchaser of its tobacco—as well as of its second major export, tea—is the United Kingdom. Sugar and cotton are the country's other major exports. Diesel fuel and petroleum, fertilizers, consumer goods, machinery and transport equipment, and medical supplies are the main imports. South Africa, Japan, the United States, Germany, and The Netherlands are Malaŵi's other major trading partners.

There are two commercial banks—the National Bank of Malaŵi and the Commercial Bank of Malaŵi. The Reserve Bank of Malaŵi is the central bank of the country. Other financial institutions include the Post Office Savings Bank, the New Building Society, and finance houses. Among the several insurance companies, only one is locally based.

Soon after independence, the government developed an economic policy that was stringently anti-inflationary, arising from the need to reduce the deficit in public expenditure and to maintain the level of foreign-exchange reserves. In budgetary policies maximum restraint, consistent with development needs and planned reduction of grants-in-aid from the United Kingdom, was exercised.

The main emphasis continues to be directed toward agricultural export production and the completion of investment projects, while at the same time maintaining a favourable balance of trade. Malaŵi's development strategy emphasizes concern for the public sector only insofar as it does not interfere with the private sector. Developmental priority is given to transport, agriculture, education, and housing.

Tax structure

A small annual tax is payable by all men over 18 years of age unless they are liable to other taxes. Employees pay an income tax. Local companies pay taxes at a fixed rate of chargeable income, and companies incorporated outside Malaŵi pay a small additional tax.

There is no sizable industrial labour force. Some 20 trade unions and employer associations are connected with such enterprises as the tea plantations, the building and construction industry, road transport, and railways. The Ministry of Labour plays a significant role in maintaining good relations between employers and employees.

Transportation. Malaŵi has road connections to Chipata on the Zambian border; to Harare, Zimbabwe, via Mwanza and Tete; and to several points on the Mozambique border. The backbone of the road system is represented by a road running from Blantyre in the south to Lilongwe in the west. A lakeshore highway runs roughly parallel to the inland highway from Mangochi to Karonga.

Of Malaŵi's two railway links to the sea, the first stretches more than 570 miles from Lilongwe eastward to the port of Beira on the Mozambique coast; an extension from Lilongwe to Mchinji, on the Zambia border, was completed in 1980. The second railroad joins the Salima-Blantyre line at Nkaya Junction to the south of Balaka and travels due east to link with the Mozambique Railways system at Cuamba, whence it continues to the port of Nacala. Increased guerrilla activity in Mozambique after 1981, including attacks on these rail lines, forced Malaŵi to seek

alternative, much longer routes to the sea, first through South Africa and then through Tanzania, adding substantially to its freight transport costs. By the early 1990s, traffic had again resumed slowly through Mozambique as Malawian and Mozambican troops were more able to suppress rebel attacks.

Of the rivers, only the Shire is partially navigable, all other streams being broken by rapids and cataracts. Lake Malawi has long been used as a means of inexpensive transportation. A passenger and cargo service that operates on the lake is linked to the Chipoka railway junction about 17 miles south of Salima. The main ports on the lake are Monkey Bay, Nkhotakota, Nkhata Bay, and Likoma Island.

Air Malawi, the national airline, operates services from the main airport at Chileka, 11 miles from Blantyre, to several foreign countries and neighbouring African capitals.

Administration and social conditions. *Government.* Under the republican constitution of Malawi promulgated in July 1966, the government is composed of a president, who is head of state and government, and the National Assembly. The cabinet is appointed by the president. The original number of 50 elected members of the assembly was raised to 60 in 1969, 87 in 1979, 112 in 1987, and 141 in 1992. In addition, the president can appoint no more than 10 nominated members.

The country is divided into 24 administrative districts. The local government system consists of district councils, the city councils of Blantyre and Lilongwe, the municipality of Zomba, and seven town councils.

Malawi was a de facto one-party state from August 1961, when the first general elections were held, until 1966, when the constitution formally recognized the Malawi Congress Party—led by President H. Kamuzu Banda—as the sole political organization. According to the constitution, elections for the presidency and the assembly are to be held every five years; general elections, however, have been held only in 1971, 1978, and 1992. The presidential candidate is nominated by an electoral college composed of party officials at the national, regional, and district levels; the League of Malawi Women; the League of Malawi Youth; members of Parliament; recognized chiefs; and all chairmen of district councils. In 1971 Banda was elected president for life. Candidates for the National Assembly may stand for election only after their nomination by the district party conferences.

Elections

Justice. The judiciary is based upon the system prevailing in the British colonial era and Malawi traditional law. It consists of a Supreme Court of Appeal, a High Court, magistrates' courts, and traditional courts. Since 1969, criminal cases involving witchcraft or local superstition, for which the death penalty can be imposed, have been tried in the traditional courts instead of the High Court. The minister of justice has the power to direct a particular case or group of cases to a particular court; cases tried in the traditional courts can be appealed to the National Traditional Court of Appeal.

Health and welfare. Health facilities include two central hospitals, district hospitals, rural clinics, Zomba Mental Hospital, and Kochira Leprosarium. Common diseases include malaria, schistosomiasis, and trachoma. Malawi is also affected by a relatively high incidence of AIDS (acquired immune deficiency syndrome), particularly in the urban areas, which threatens to tax the country's overburdened health-care system even more. Malawi has the highest infant mortality rate in southern African and one of the highest population-to-physician ratios in all of sub-Saharan Africa.

An acute shortage of housing has existed for several years in urban areas. The Malawi Housing Corporation has launched several projects to build houses and develop traditional housing areas.

Education. Elementary education is not compulsory, and only about one-half of all eligible children attend primary school. Despite this low proportion, Malawi's primary schools feature one of the highest student-teacher ratios in Africa. Postprimary education comprises a four-year secondary-school course that can lead to a university education. The Malawi Correspondence College is available to students unable to attend regular secondary school. There are also institutions for teacher training and for technical and vocational training. The Kamuzu Academy at Mtunthama is a secondary school for gifted children. The University of Malawi, founded in 1964, has four constituent colleges.

Cultural life. Though under the impact of modernization, Malawi's traditional culture is characterized by continuity as well as change, and the traditional life of the village has remained largely intact. One of the most distinctive features of Malawi culture is the enormous variety of traditional songs and dances that use the drum as the major musical instrument. Among the most notable of these dances are *ingoma* and *gule wa mkulu* for men and *chimtali* and *visekese* for women. There are various traditional arts and crafts, including sculpture in wood and ivory. There are two museums—the Museum of Malawi in Blantyre and a smaller one in Mangochi. While various cultural activities are organized by the Ministry of Youth and Culture, the University of Malawi Travelling Theatre, and other groups in Blantyre, the radio from Zomba and Lilongwe has proved to be the most effective means of bringing traditional and modern plays to the rural population. (Z.D.K./Ed.)

Traditional dances

For statistical data on the land and people of Malawi, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The paleontological record of human cultural artifacts in Malawi dates back more than 50,000 years, although known fossil remains of early *Homo sapiens* belong to the period between 8000 and 2000 BC. These prehistoric forebears have affinities to the San (Bushmen) of southern Africa and were probably ancestral to the Twa and Fula, whom Bantu-speaking peoples claimed to have found when they invaded the Malawi region between the 1st and 4th centuries AD. From then to about AD 1200, Bantu settlement patterns spread, as did ironworking and the slash-and-burn method of cultivation. The identity of these early Bantu-speaking inhabitants is uncertain. According to oral tradition, names such as Kalimanjira, Katanga, and Zimba are associated with them.

With the arrival of another wave of Bantu-speaking peoples between the 13th and 15th centuries AD, the recorded history of the Malawi region began. These peoples migrated into the region from the north, and they interacted with and assimilated the earlier pre-Bantu and Bantu inhabitants. The descendants of these peoples maintained a rich oral history, and, from 1500, written records were kept in Portuguese and English.

Among the notable accomplishments of the last group of Bantu immigrants was the creation of political states or the introduction of centralized systems of government. They established the Maravi Confederacy about 1480. During the 16th century, the confederacy encompassed the greater part of what is now central and southern Malawi, and, at the height of its influence in the 17th century, its system of government affected peoples in the adjacent areas of modern Zambia and Mozambique. North of the Maravi territory, the Ngonde founded a kingdom about 1600. In the 18th century, a group of immigrants from the eastern side of Lake Malawi created the Chikulamayembe state to the south of the Ngonde.

Early political units

The precolonial period witnessed other important developments. In the 18th and 19th centuries, better and more productive agricultural practices were adopted. In some parts of the Malawi region, shifting cultivation of indigenous varieties of millet and sorghum began to give way to more intensive cultivation of crops with a higher carbohydrate content, such as corn, cassava, and rice.

The independent growth of indigenous governments and improved economic systems was severely disturbed by the development of the slave trade in the late 18th century and by the arrival of foreign intruders in the late 19th century. The slave trade in Malawi increased dramatically between 1790 and 1860 because of the growing demand for slaves on Africa's east coast. Swahili-speaking people from the east coast and the Ngoni and Yao peoples

entered the Malawi region between 1830 and 1860 as traders or as armed refugees fleeing the Zulu states to the south. All of them eventually created spheres of influence within which they became the dominant ruling class. The Swahili speakers and the Yao also played a major role in the slave trade.

Islām and Christianity

Islām spread into Malawi from the east coast. It was first introduced at Nkhotakota by the ruling Swahili-speaking slave traders, the Jumbe, in the 1860s. Traders returning from the coast in the 1870s and '80s brought Islām to the Yao of the Shire Highlands. Christianity was introduced in the 1860s by David Livingstone and by other Scottish missionaries who came to Malawi after his death in 1873. Missionaries of the Dutch Reformed church of South Africa and the White Fathers of the Roman Catholic church arrived between 1880 and 1910.

Christianity owed its success to the protection given to the missionaries by the colonial government, which the British established after occupying the Malawi region in the 1880s and '90s. British colonial authority was welcomed by the missionaries and some African societies but was strongly resisted by the Yao, Chewa, and others. In 1891 the British established the Nyasaland Districts Protectorate, which was called the British Central Africa Protectorate from 1893 and Nyasaland from 1907.

Under the colonial regime, roads and railways were built, the cultivation of cash crops by European settlers was introduced, and inhumane practices were suppressed. On the other hand, the colonial administration did little to enhance the welfare of the African majority because of its commitment to the interests of the European settlers. It failed to develop African agriculture, and many able-bodied men migrated to neighbouring countries to seek employment. Furthermore, between 1951 and 1953 the colonial government decided to join the colonies of Southern and Northern Rhodesia and Nyasaland in the Federation of Rhodesia and Nyasaland, against bitter opposition from their African inhabitants.

These negative features of colonial rule prompted the rise of a nationalist movement. From its humble beginnings during the period between the world wars, African nationalism gathered momentum in the early 1950s. Of special impetus was the imposition of the federation, which nationalists feared as an extension of colonial power. The full force of nationalism as an instrument of change became evident after 1958 under the leadership of Hastings Kamuzu Banda. The federation was dissolved in 1963, and Malawi became independent as a member of the Commonwealth of Nations on July 6, 1964.

(Z.D.K./K.M.G.P./Ed.)

Soon after independence, a serious dispute arose between Banda, the prime minister, and other ministers. In September 1964 three ministers were dismissed and three others resigned in sympathy. Henry Chipembere, one of these ministers, escaped from house arrest and defied attempts to recapture him, becoming the focus for antigovernment opinion until his death in 1975. On July 6, 1966, Malawi became a republic, and Banda was elected president.

Banda's presidency

Malawi's 1966 constitution established a one-party state under the Malawi Congress Party (MCP), which in turn was controlled by Banda, who consistently suppressed any opposition. The conservative MCP concentrated its attention on economic development. In foreign policy Banda adopted a position that was in opposition to most countries in sub-Saharan Africa: he established friendly trading relations with the Republic of South Africa and appealed to other African leaders to be more realistic in their attitude toward racial problems. Securing access to transportation routes to coastal ports was a major factor in Malawi's regional policy.

Although Banda was made president for life in 1971, political agitation forced him to accept multiparty elections for the National Assembly and the presidency. In 1994 presidential elections were won by Bakili Muluzi, who led the United Democratic Front (UDF). Muluzi released or lowered the sentences of various political prisoners and focused on food production, alleviating poverty, and ending governmental corruption. Banda and others were arrested for the murders of several government officials in 1983, but they were acquitted in 1995. Before his death in 1997, Banda apologized to Malawians for any acts of brutality his regime had committed. Muluzi was reelected in 1999, and the UDF continued to dominate the government. Despite his intentions, Muluzi was not entirely successful in eliminating corruption and poverty, and food production continued to fall short of demand. The country also faced a difficult challenge as the number of AIDS cases continued to grow at the beginning of the 21st century. (Ed.)

For later developments in the history of Malawi, see the BRITANNICA BOOK OF THE YEAR.

Mozambique

The Republic of Mozambique (República de Moçambique), formerly the People's Republic of Mozambique, stretches along the Indian Ocean coast of southeastern Africa from Cape Delgado at latitude 10°27′ S to latitude 26°52′ S. Its westernmost border at the Aruângua (Luangwa) River reaches longitude 30°31′ E, and the easternmost point, 110 miles (175 kilometres) east of Nampula on the Indian Ocean coast, is at longitude 40°51′ E, but most of the country's 313,661 square miles (812,379 square kilometres) lies between longitudes 32° and 40° E. It is bordered to the south and southwest by the Republic of South Africa and Swaziland, to the west by Zimbabwe, to the northwest by Zambia, Lake Nyasa (Niassa), and Malawi, and to the north by Tanzania. The Mozambique Channel separates it from Madagascar to the east. The capital city of Maputo (formerly Lourenço Marques) is in the nation's southernmost province.

Mozambique's extensive coastline (1,563 miles) features some of Africa's best natural harbours, a fact that contributes to the nation's important transportation and communication role in the region. The massive Zambezi River dominates the central area and provides sufficient hydroelectric potential to make Mozambique the region's powerhouse.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Lowlands dominate the southern provinces, narrowing to a mere coastal plain north of the cleft that the Zambezi River cuts through the country's midsection. The Zambezi valley, the lower section of which is a part of the Eastern (Great) Rift Valley, is the country's most dramatic geographic feature. Throughout the country the land rises gently to the west. In the centre and north it slopes steadily into the high plains, and ultimately the mountainous regions of the northwest border Malawi and Zambia. Four of Mozambique's five highland regions straddle the border areas in the west and northwest: the Chimoio Plateau on the border with Zimbabwe, the Maravia highlands bordering Zambia, and the Angónia highlands and Lichinga Plateau, which lie, respectively, west and east of Malawi's protrusion into Mozambique. Mount Binga, the country's highest elevation at 7,992 feet (2,436 metres), is part of the Chimoio highlands. The 7,936-foot (2,419-metre) peak at Mount Namúli dominates the Mozambican highland, which constitutes much of the northern interior.

Highland regions

Drainage. Mozambique's many rivers hold the promise of irrigation for agriculture and hydroelectric power for the entire region. The Rovuma River defines most of Mozambique's northern border with Tanzania. The Zambezi River and its tributaries dominate the central region, and the Maputo River forms part of the southernmost boundary with Swaziland and South Africa. Rivers—including the Lúrio, Ligonha, Save, Changane, and Komati (Incomati)—also define many of the country's local political boundaries. Other important drainage systems include the Messalo River in the north and the Púngoè, Revuè, and Búzi rivers, which enter the Mozambique Channel together just south of the port of Beira.

The massive Zambezi waterway clearly dominates the nation's drainage and its hydroelectric strategies. It flows 509 miles through the country and drains more than 87,000 square miles of the central region. The Rovuma, Lúrio, Save, and Messalo systems follow in size. Mozambique shares the borders of Lakes Nyasa, Chiuta, and

Chilwa with Malawi, but aside from these border lakes and the lakes created by the country's hydroelectric dam systems—particularly the extensive system created by the Cabora Bassa Dam at Songo on the Zambezi—the country has no important lakes.

Soils. Africa's ancient basement complex of granite underlies most of northern and west-central Mozambique, whereas the soils of the southern and east-central regions are sedimentary. Mozambique's soils are diverse in quality and type, but the central and northern provinces enjoy generally more fertile, water-retentive soils than does the south, where sandy, infertile soils prevail. South of Beira fertility is largely limited to alluvial soils in the valleys of the Save, Limpopo, Komati, Umbelúzi, and Maputo rivers. Several pockets of fertile but heavy soil occur southwest of Inhambane. The central region enjoys its broadest expanse of rich alluvial soils along the Zambezi delta. The northern soils have higher red clay content and range from infertile to quite fertile. Their water-retentive qualities allow agricultural potential to extend beyond the river valleys.

Monsoon influence

Climate. Mozambique lies largely within the tropics, and much of the coastline is subject to the regular seasonal influence of the Indian Ocean monsoon rains. The monsoon influence is strongest in the northeast but is modified somewhat in the south by the island barriers of Madagascar, the Comoros, and the Seychelles. With the exception of highland areas on the northern and western borders and around Gurue (east of the Malawi protrusion into Mozambique), all locations where altitude modifies both temperature and humidity, the climate is seasonal and tropical. Daily temperatures throughout the country average in the mid- to upper 70s F (lower to mid-20s C), with the highest temperatures occurring between October and February and the lowest in June and July. Uncomfortably warm average daily temperatures, in the upper 80s F (low 30s C), are normal only in the upper Zambezi valley and along the northeastern coast, while cool temperatures, in the 60s F (10s C), occur year-round only in the mountainous areas on the western borders.

Humidity and rainfall vary a great deal throughout the country. Again, the sharpest difference is between north and south. The entire region north of the Zambezi and east of the Shire River valley is humid and warm, as is the coastal plain in the south. The southern interior and most of the Zambezi valley west of the Shire is quite dry, and the south-central area around Pafuri is even considered semiarid, receiving only about 2.5 inches (60 millimetres) of rainfall per month in the wet season from November to February and almost none in the dry season between April and October. In the south, to the west of the coastal plain, average annual rainfall is only about 24 inches. Rainfall in the central region east of the Shire River and throughout the north is much higher, between 40 and 70 inches, with the highest rainfall averaging more than 70 inches in the highlands and in coastal pockets around Beira and Quelimane. In the Zambezi valley west of the Shire, however, average rainfall dips to the 24- to 32-inch levels typical of the south. As the annual rainfall figures suggest, southern and west-central Mozambique are subject to drought crises. The drought of 1992 was the worst since records have been kept.

Plant and animal life. Although Mozambique retains some dense forestlands in the north-central interior and on the Chimoio Plateau, most of the north and east-central areas are open forest. In the south the open forest of the east becomes brush and savanna grassland to the west. The largest forest reserves are on the Chimoio Plateau west and southwest of Beira and in the northern interior south of the Lúrio River. Mozambique maintains four national parks in the central and southern areas—Gorongosa, Zinave, Bazaruto, and Banhine.

Wildlife populations include water buffalo, elephant, warthog, leopard, baboon, giraffe, zebra, antelope, lion, and numerous species of ungulate and cat. Crocodiles and hippopotamuses are still found in slow-moving waterways. Snakes—including impressive pythons and dangerous puff adders, cobras, and vipers—are found throughout the territory. Flamingos, cranes, storks, herons, pelicans, ibis, and other tropical water birds exist throughout Mozambique but are more numerous in the moister areas of the northeast. Scavengers include crows, vultures, and buzzards, and game birds include guinea fowl, partridge, quail, and a range of geese and ducks. Game reservations and national hunting areas are located largely in the central and southern areas, with the exception of the important Niassa reserve on the Tanzanian border and the Gilé reserve southwest of Nampula. The largest game areas are just south of the Zambezi bordering the Chimoio highlands. The nation's five hunting reservations are Niassa, Gilé, Marromeu, Pomene, and Maputo.

Birdlife

Settlement patterns. Since Mozambique is an overwhelmingly agricultural country, with more than 80 percent of the labour force involved in farming, settlement patterns reflect the country's agricultural potential. Areas with the best soils and climates are the most populous. The Lúrio and Ligonha river valleys in the northeast, as well as the coastal plain between them, are densely populated, as are the lower reaches of the Limpopo valley and the limited areas of deep rich soil south and west of Inhambane. In most rural areas settlements reflect family arrangements and are dispersed. In drier areas settlement patterns are shaped by efforts to combine agriculture with pastoralism. Settlements are separated by expanses of grazing area. People in small settlements typically plant several crops in diverse and specific environments to minimize the danger of famine in the case of flood, drought, pests, or other natural stress.

The Portuguese colonial state developed rural settlement schemes during the late colonial era, and shortly after independence in 1975 the national government strongly promoted communal village and state farm projects, all of which fostered denser rural settlement, particularly in the south. In most cases, however, such schemes proved largely unsatisfactory, and more dispersed settlement patterns tended to reemerge when government policy and military security permitted. Although dispersed settlement makes it more difficult for the state to provide security and community services, it is preferred by farmers.

Major cities

Maputo is the country's principal urban settlement, followed by Beira, Nampula, Nacala, Quelimane, Pemba, and Chimoio. Most of these urban settlements are port, transportation, and communications centres, which grew in response to the service needs of Mozambique's western neighbours. The development of Nampula, Nacala, and Chimoio dates from the Portuguese colonial state's efforts to decentralize economic and administrative infrastructure as part of its counterinsurgency strategy of the late 1960s and '70s. Prior to that period investment in Mozambique was strongly focused on Maputo. Only Maputo and Beira have substantial foreign communities.

The people. *Ethnic and linguistic composition.* Although Portuguese is the official language, the vast majority of Mozambicans speak languages of the Niger-Congo group, the so-called Bantu languages, which dominate central and southern Africa. Within that group, Makua-Lomwe, Tsonga, and Shona are the most widespread languages, but the country has great linguistic and cultural variety. Language groups in the Zambezi valley are quite diverse and include Sena, Lomwe, and Chuabo. Mozambicans share many languages with their neighbours, including Swahili, Yao, and Makonde with Tanzanians; Nyanja and Chewa with Malawians; Shona with Zimbabweans; and Shangaan with people of the northeastern Transvaal in South Africa. The Swahili speakers of Mozambique's northern coast have an Islāmic heritage in common with the coastal populations of eastern Africa as far north as Mogadishu, Somalia. Similarly, small groups in the far south and throughout the country share Nguni languages with South African and Zimbabwean peoples as a result of the important population movements of the early 19th century. Groups speaking European and Asian languages are largely limited to the port cities of Maputo, Beira, Quelimane, Nacala, and Pemba.

In terms of cultural organization, the Zambezi valley again provides Mozambique's key marker, roughly dividing groups that trace their heritage according to principles of matrilineality to the north and groups that order them-

selves along patrilineal lines to the south. In matrilineal groups authority rests in the senior male of the extended family traced through the female line, whereas in patrilineal groups the senior male is identified through the male line. Throughout the 20th century, however, there was a tendency for once-matrilineal groups to adopt patrilineality and virilocal settlement—with new families settling in a household of the husband's lineage rather than the wife's.

Of the more than 30 languages spoken in Mozambique, Bantu languages—a subgroup of the Benue-Congo branch of the Niger-Congo language family—compose the vast majority. One of these, Makua, is spoken by almost one-third of the population. Tsonga (Shangaan, Changana), which is spoken south of the Sabi River, has more than two million speakers, and Lomwe, Sena, and Chuabo more than one million speakers each. Both Lomwe and Chuabo are closely related to Makua. Lomwe is spoken in the northeast and central regions, and Sena mainly in the northwest. Chuabo is the language of the central coast, and Swahili is the language typically of the coastal strip north of the Lúrio River. Chopi, Makonde, Marendje, Nsenga, Shona, Tswa, Yao, Chewa (Cewa), and Nyanja languages are also spoken.

Although Portuguese is the main language of only a small fraction of the people, it is spoken as a lingua franca by some two-fifths of the country's inhabitants. Portuguese speakers are strongly concentrated in Quelimane and the capital of Maputo.

Religions. Prior to independence in 1975, almost one-third of the population was nominally Christian, and a small number were Muslim. Christian missions were active throughout the country during the colonial era, and after 1926 the Roman Catholic church was given government subsidies and a privileged position with respect to its educational and evangelical activities among the African population. Although the Portuguese were generally suspicious of Protestants, Protestant missions—including Presbyterian, Free Methodist, African Methodist Episcopal, Methodist Episcopal, Anglican, and Congregationalist missions—remained active. A variety of African independent churches developed, but because of official disdain for their activities they were unlikely to register publicly.

Mission activities

After independence the government, led by the Mozambique Liberation Front (Frente de Libertação de Moçambique; Frelimo), presented conflicting messages regarding religion. Although it confirmed a policy of open and free religious affiliation, Frelimo actively persecuted the country's more than 20,000 Jehovah's Witnesses, and its overall political and ideological emphasis discouraged religious expression and organization. By the end of the 1980s, however, Frelimo had changed its approach, and religious organizations began to reemerge as an important popular force.

Almost half of the people now follow traditional religions, while fewer than one-fourth adhere to some form of Christianity, and almost three-tenths are Muslims. Although Islāmic communities are found in most of Mozambique's cities, Muslims constitute the majority in only the northern coastal region between the Lúrio and Rovuma rivers.

Demographic trends. Mozambique's rate of population growth, though high by world standards, is average for southern Africa. The country's demographic profile reflects the poverty, health, and political situation of the majority population of southern Africa. Malnutrition, insufficient access to clean water supplies, insufficient sanitation facilities, and intractable, incurable, and endemic diseases all take an enormous toll. Mozambique's rate of infant mortality is among the highest in the world. Children have less than a one-in-four chance of surviving to age 5, and average life expectancy is under 50 years. As in most African nations, Mozambique's population is young—more than 40 percent is under the age of 15 and 70 percent is under 30.

Population movement across Mozambique's borders has been facilitated in many instances by shared language and culture. During the colonial era Mozambicans worked in neighbouring countries as contract labourers and independent migrant workers, particularly in the mining areas of South Africa and in the farms and cities of Southern Rhodesia (now Zimbabwe). After the coup d'état in Portugal (April 24, 1974) that signaled the end of Portuguese colonial rule in Africa, the Portuguese population in Mozambique plummeted from a high of about 250,000 to fewer than 10,000. Within five years of independence, social and economic destabilization by the antigovernment forces of the Mozambique National Resistance (Resistência Nacional Moçambicana; Renamo) brought about extensive population dislocation in the Mozambican countryside. Disruption of rural production and distribution, natural disaster, and counterproductive government agricultural and commercial policies fueled a generalized economic collapse. By the mid-1980s almost one-third of the nation's population had left their fields and herds and fled to refugee settlements around the major cities and in neighbouring countries.

Population movement

The economy. Mozambique's predominantly rain-fed agricultural economy is based on family production and hoe technology. During the 20th century, plantation production of market crops displaced family agriculture in some of the most fertile areas. The colonial economy was characterized by private monopolies, central planning, and state marketing of key products, all designed to promote capital accumulation by the state, Portuguese settlers, and Portuguese-based commerce and industry. Colonial policy had also excluded most Mozambicans from highly skilled and managerial positions until the years immediately preceding independence. The Frelimo government tried to redirect patterns of accumulation and development by nationalizing key properties, promoting African education and training, and breaking up the Portuguese and Asian hold on commercial distribution. Despite Frelimo's public stand against ethnic discrimination, Portuguese settlers and Asian traders—threatened by the government's economic policies—left by the thousands. Settlers anticipating nationalization abandoned their properties, adding by default to the proportion of the national economy that the state controlled. Large-scale state-run farms and communal and cooperative farming replaced the settler and company plantations. Frelimo's agricultural undertakings proved unproductive and unmanageable, and, in combination with the flight of Asian merchants, much of the nation's agricultural production, commercial, and distribution sectors collapsed. The state ultimately reoriented economic policy in accordance with structural adjustment plans imposed by the International Monetary Fund (IMF), which emphasized decentralization and assisted family farming in an attempt to salvage and rebuild the economy.

Although agriculture is the most common economic activity, migrant labour remittances from South Africa, tourism, and the nation's port and railway sector have historically been equally important sources of foreign-exchange revenues. While all these sectors declined severely during the 1980s and early '90s because of civil unrest, they began to rebound after the 1992 peace accord, and resource exploitation also increased. The more open political situation also allowed numerous trade unions to develop, many of which participated in the Organization of Mozambican Workers (Organização de Trabalhadores Moçambicanos), a group that has openly criticized the free-market policies of the government.

In both the colonial and postcolonial eras state regulation of the economy was common, but the methods contrasted sharply. Particularly during the 20th century, Portuguese investors were allowed to develop important components of the economy without competition. Only in the post-World War II era were economic decentralization and competitive markets allowed to develop apace. After independence the state was weak, and its efforts to develop and expand the economy through central planning were ineffective and counterproductive.

Economic regulation

Resources. Although much of Mozambique has not been extensively surveyed for minerals and petrochemical resources, from what is known the country has impressive potential energy resources. The Tete highlands in the west-

central region has six billion tons of known bituminous coal reserves, and the Cabora Bassa Dam (Hidroeléctrica de Cabora Bassa) was completed in 1974 to tap the hydroelectric potential of the Zambezi River. Although continuing exploration for oil has been disappointing, large commercially viable natural gas fields were discovered at Panda, north of Inhambane, and are in the early stages of development.

Key known mineral resources include high-quality iron ore reserves and what may be the world's largest reserves of the rare and important mineral tantalite. Tantalite is essential in the global electronics industry and is used to produce top-grade steels. Mozambique's small gold industry is also expected to develop rapidly.

Water resources could potentially compensate for the mixed soil endowment. Mozambique's major river systems provide alluvial deposits and offer both hydroelectric and irrigation potential. The country's forest resources were briefly exploited for both export and local building material during the 1960s and '70s, but significant hardwood forest reserves have survived deforestation for fuel.

Mozambique's offshore waters contain tuna, mackerel, sardines, and anchovies but are best known for the shrimp (prawns) that are an important export commodity. Availability of shrimp, lobster, and shellfish have also contributed to Mozambique's attraction as a vacation beach resort for its inland neighbours. The pleasant climate, beautiful beaches, and Indian Ocean islands are important resources. Mozambique's natural resources remain largely underdeveloped owing to military insecurity and lack of investment capital.

Agriculture, forestry, and fishing. Agriculture contributes about 40 percent of Mozambique's gross domestic product (GDP), but, in most rural areas since the late 1970s, agricultural output has declined steadily. Most agricultural production is undertaken with family labour to produce the two staple crops, corn (maize) and cassava, as well as beans, rice, and a variety of vegetables and oilseeds. Family labour also gathers a large part of the important cashew nut crop and produces cotton, beef, and small stock for the local market and export.

Agricultural exports, such as sugar, tea, citrus, copra, and sisal, were developed as plantation crops during the colonial era and, with few exceptions, continue to be produced on private plantations, state farms, and government-sponsored cooperatives and communal villages, although production of each crop is down from preindependence levels. Irrigation remains limited to schemes developed in the 1950s and '60s in former settler areas in the south, particularly along the Limpopo River. Beef production is important in the drier southern sections of the country. Egg, poultry, milk, and pork production passed from settler to state control after independence, but only egg and poultry production has sustained or surpassed output from the early 1970s.

Since independence the exploitation of timber has declined, though international investors are interested in Mozambique's potential to provide wood for building materials and pulp for paper industries. Today, however, the forests are principally exploited as the nation's key source of domestic fuel, firewood, and charcoal. More timber is cut than is replaced by reforestation initiatives.

Fishing is one area of the economy that rural insecurity has not severely undermined, and, since 1973, production and marketing of saltwater fish, shrimp, and shellfish have increased steadily. Shrimp constitute an increasingly important proportion of foreign-exchange earnings.

Industry. The mining industry in most of Mozambique remains either closed or well below capacity owing to the security situation. The large reserves of tantalite at Murrua had been developed by East German and Soviet interests until the collapse of the Eastern-bloc governments. Investors from several areas, including South Africa, have expressed an interest in pursuing the tantalite deposits.

High-quality iron ore, bauxite, graphite, copper, marble, garnets, asbestos, bentonite, limestone, and sea salt are all mined and quarried in Mozambique. Western European, American, and Japanese investors have all expressed interest in the mining sector. Frelimo reforms regarding international investment and the collapse of Eastern-bloc influence in the region have fueled Western interest.

In 1975 Mozambique's industrial and manufacturing sector was typical of much of colonial Africa, based largely on minimally processed raw materials for export (shrimp, tea, sugar, cashews, citrus, copra, coal, and cotton) combined with processing and light manufacturing for local consumption (milled grains, beer, tobacco, soap, building materials, vegetable oils, and, to a lesser extent, textiles and shoes). A sharp increase in local manufacturing capacity occurred from the late 1950s to the early 1970s in response to consumer demand from the burgeoning urban settler population. The limited development of heavier industry was linked to trade, service, and transportation agreements with the neighbouring countries, such as the metalworking and heavy railway equipment sector. The enormous expansion of cement production was due to the construction of the Cabora Bassa Dam and urban housing fueled by the settler influx.

Despite the exodus of much of the country's skilled labour and managerial class at independence, industrial production increased modestly until the early 1980s, after which point manufacturing and construction ground to a virtual halt. Within 15 years of independence, industrial production had declined by almost two-thirds. War interrupted the supply of raw materials from within the country, and the nation's soaring debt and diminishing exports combined to strangle Mozambique's ability to import spare parts and other inputs necessary to sustain production. Sabotage and power outages exacerbated the situation. By the late 1980s and early 1990s, debt restructuring, negotiations with international creditors, and a somewhat improved business climate for private investors alleviated capital loss and shortages of foreign exchange, enabling businesses to purchase spare parts and capital goods.

On the strength of its resources, Mozambique should be southern Africa's most important energy producer and exporter, but continuing military insecurity has undercut production in all but the case of imported crude oil at the Matola refinery upriver from Maputo. The centrepiece of Mozambique's energy potential is the Cabora Bassa Dam on the upper Zambezi River. It was financed and constructed by an international consortium at the close of the colonial era. It was designed in cooperation with South Africa's national power company, Eskom, to produce electricity for South Africa, not for Mozambique. Upon completion in 1974, only 150 megawatts of the more than 2,000-megawatt capacity were to be consumed in Mozambique; the rest were to travel via the world's longest (870 miles) direct-current line to the Apollo power station in South Africa's industrial heartland. The Portuguese hoped South Africa's interest in the dam would guarantee its support of Portugal during the independence struggle then raging in Mozambique. Since independence, however, Mozambique has increased the national share of ownership in Cabora Bassa, and by the year 2014 it will hold full ownership. The destruction of hundreds of current-line pylons during the postindependence conflict eclipsed the power flow before it was effectively begun. The entire national electrical grid has been targeted by Renamo, frustrating energy-development efforts by forcing cities to build and improve self-contained facilities that do not make the best use of the nation's hydroelectric potential. Cabora Bassa Hydroelectric and two privately run dams on the Revuè River are the principal hydroelectric stations. About 90 percent of Mozambique's electricity is still generated by variously fueled thermal power plants.

The Cabora Bassa Dam

Since the 1980s, Western investors have been attracted to the tourist sector. British and South African investors had important interests in this sector before independence and have recently renewed them.

Finance. In 1978 Mozambique nationalized most of the nation's banking assets. The state consolidated the banking sector into two institutions, Banco de Moçambique, the bank of issue, and Banco Popular de Desenvolvimento (Peoples' Development Bank). To accommodate the shortage of foreign exchange to fund development and eventually to import basic necessities such as food, Mozambique borrowed on the international market to

the point that by the 1990s debt service consumed one-quarter of the nation's annual foreign-exchange earnings. In 1983 Mozambique agreed to join the World Bank and International Monetary Fund and to adopt structural adjustment and national economic recovery programs as a condition for new loans and grants. Private and international investment were encouraged in the new economic climate. International enthusiasm for continued investment hinges on a sustained peace accord between Frelimo and Renamo.

Trade. The most important exports by value include cashew nuts, shrimp, lint cotton, sugar, citrus, copra, timber, tea, and coal. They are sold mostly to the Organization for Economic Cooperation and Development (OECD) nations, among which Spain, the United States, and Japan are the leading consumers. Japan has sharply increased its purchase of Mozambique's exports, while sales to former Eastern-bloc countries have declined.

Trading partners

Food is Mozambique's single largest import by value, followed by equipment, spare parts, and crude oil. Just under half those imports comes from OECD nations, with the United States, Portugal, and Italy providing the largest share. About a third comes from African nations, principally South Africa. The import sector reflects the same pattern of Japanese ascendency and former Eastern-bloc decline revealed in exports. Despite the substantial increase in exports by value since the mid-1980s, imports have vastly outstripped those gains, more than doubling the negative balance of trade. Such patterns are likely to continue if Mozambique is unable to secure rural areas so that refugees can return to agricultural production.

Transportation. Mozambique's transportation sector reflects the nation's historical development in relation to its neighbours. The national road, railway, and port sectors were originally developed by the state and by chartered companies primarily to service the trade and transport needs of Mozambique's western neighbours. East-to-west rail and road systems linking Mozambique's excellent ports with the key industrial and mining areas of South Africa, Zimbabwe, and Malawi are well developed, whereas road and rail systems facilitating north-to-south intra-Mozambique transportation are virtually nonexistent in the case of railroads and minimal with regard to permanent roads. Most of the existing network of internal connecting roads and airstrips in the northern and central areas were developed during the 1960s and '70s as part of Portugal's counterinsurgency strategy. Air transportation has continued to be developed owing to the insecurity of rural roadways.

Mozambique's potential as a transport centre for the interior is on par with its energy capabilities. Its three international ports—Maputo, Beira, and Nacala—are among the best on the continent. Maputo is linked by rail to Swaziland, South Africa, and Zimbabwe. Beira services Zimbabwe by road and rail through what has come to be called the Beira Corridor. It is also linked by rail to Malawi and to the Moatize coal mines near Tete. Nacala's railway extends through Nampula and Cuamba to Malawi, with a leg north from Cuamba to Lichinga.

Ten small local ports dot the coastline from Tambuzi in the north to Inhambane in the south, but only half of those have even limited rail-line links to their hinterlands. The port and railway complex at Maputo was developed at the end of the 19th century in response to the developing gold- and coal-mining industries of Johannesburg and the Transvaal region of South Africa. The colonial state managed to link the mining industry's access to Mozambican contract labour with a commitment to export a substantial fixed portion of the region's mineral exports through Maputo, thus guaranteeing service and customs revenues for the port. Subsequent rail lines linked Maputo with Swaziland and ultimately with Zimbabwe's Gweru (Gwelo) mining area (1955). Maputo has two miles of deep-water wharf and specialized facilities for citrus cold storage, sugar, molasses, coal, and steel. In the 1950s the port and railway administration greatly expanded a specialized industrial port for petroleum, timber, and minerals (iron ore and chromite) 1.5 miles upriver at Matola.

The port of Maputo

The rail links from Beira to Zimbabwe and Malawi were originally developed by the Mozambique Company and taken over by the Portuguese colonial government in 1947. Traffic on the Beira line is dominated by Zimbabwean minerals (chromite, copper, and asbestos) and agricultural products (corn and sugar). With the independence of Zimbabwe in 1980, international investors—particularly British firms—took a renewed interest in rehabilitating and upgrading the Beira Corridor, but the line from Beira to Malawi has been closed because of the war. Beira has about a mile of wharf and special coaling, citrus cold storage, molasses, and tallow facilities.

Nacala, with the nation's best natural harbour and newest facilities, is well placed to serve agricultural development in the north. Malawi developed a new railway line to connect with Nacala's port and railway via Zomba, but attacks on the line forced its closure in the mid-1980s.

Ferry service is available along the lower Zambezi at both Luabo and Marromeu and above Cabora Bassa between Chicoa and Zumbo, which lies near the point where Mozambique meets both Zambia and Zimbabwe. Ferries also serve Lake Nyasa, with Mozambican ports at Metangula and Meponda.

Private aircraft were the first to fly regularly in Mozambique, but after World War II Portugal's national airline, Transportes Aéreos de Portugal (TAP; Portuguese Airlines), opened a route between Beira and Maputo. Eventually colonial Mozambique developed its own airline, Direcção de Exploração dos Transportes Aéreos (DETA; Directorate of Air Transport Exploitation). In 1980 DETA was replaced by Linhas Aéreas de Moçambique (LAM; Mozambican Air Lines), which remains the national carrier. Mozambique's small fleet services national and international routes. LAM has regular scheduled flights to neighbouring capital cities as well as flights to several European capitals and to Rio de Janeiro.

Air services

Expansion of national air traffic is a direct outgrowth of rural insecurity. Airline passenger traffic has developed in inverse proportion to that of the railways and roads,

Michel Huet—HOA-QUI

The port and railway complex at Maputo, Mozambique.

increasing steadily from the late 1970s as road and railway passage declined with the threat of ambush. The country is better served for north-south domestic travel by the airlines than by the more east-west biased road or rail system. Mozambique has 16 airports, of which 2 offer international service.

Administration and social conditions. *Government.* Mozambique is in transition from a single-party state with a strong commitment to socialism to a multiparty system of still uncertain orientation and from a centrally planned, largely state-owned economy to a mixed economy encouraging private ownership and international investment. Both sets of changes will have important implications for Mozambique's government, political processes, and social conditions.

Frelimo

Frelimo led the armed insurgency against Portuguese colonial rule and came to power precipitously in 1975 after brief negotiations with the Armed Forces Movement, which had toppled the government in Lisbon in 1974. Frelimo developed in the early 1960s as one of many socialist-oriented guerrilla groups seeking to overturn colonialism and white minority rule in southern Africa and saw Mozambique's independence as a component of the regional struggle against white domination. That commitment necessarily involved Mozambique in the continuing struggle in Zimbabwe and South Africa.

Mozambique's government was structured by the national constitution, produced by the Central Committee of Frelimo and set forth at independence. Under the constitution Mozambique's president, who was also the president of Frelimo, headed the Council of Ministers, the elected People's Assembly (more than 200 members), and the party's Central Committee. He was also the commander in chief of the armed forces. Since membership in Frelimo was a prerequisite for any political office, the most powerful national and provincial offices tended to circulate among a fairly small group of trusted party members.

The 1975 constitution set forth the spirit of the law. While popular assemblies were in the process of articulating a specific body of legislation, colonial legislation was allowed to stand unless it was specifically judged to contradict the spirit of the new constitution. Legislation and judicial principles and practice evolved piecemeal through the work of popular assemblies and popular tribunals.

In November 1990 a new constitution introduced sweeping changes in the government. Candidates from competing parties were to be elected by universal adult suffrage and secret ballot. The president was limited to three consecutive five-year terms. The constitution established a parliament with limited ability to veto executive action. The new constitution abolished the death penalty and confirmed freedom of the press, the workers' right to strike, and the concept of habeas corpus.

The Assembly of the Republic (Assembleia de la República), the new legislature, contains from 200 to 250 members who are elected to five-year terms by universal suffrage. Among its powers are the ability to ratify the suspension of constitutional guarantees, approve the appointment of the president and deputy president of the Supreme Court, and grant amnesties and pardons.

The judicial system consists of professional judges and elected judges. Professional judges, appointed by the president in consultation with the Supreme Council of the Judiciary, have jurisdiction over matters of law while elected judges are concerned with the primary trial courts.

Mozambique is divided into 10 provinces: Cabo Delgado, Niassa, Nampula, Zambézia, Tete, Sofala, Manica, Inhambane, Gaza, and Maputo; the president appoints the governor of each province. The provinces, in turn, are divided into 112 districts and 894 localities.

Education. The Portuguese educational system was two-tiered—designed to promote rudimentary skills among the majority African population and to provide liberal and technical education for the settler population and a tiny minority of Africans. Nearly 90 percent of students enrolled in the colonial system were restricted to the rudimentary program. The state, in cooperation with the Roman Catholic church, provided public education, but private education was also available, mostly through church groups. The medium for education was Portuguese, but at independence less than 7 percent of the African population was literate in that language.

Education since independence

Private and parochial school facilities were nationalized to facilitate the unification and total overhaul of the educational system. Demand for education quickly outstripped the state's capacity. The number of students at the primary level doubled and at the secondary level tripled. Literacy programs were extended to the adult population at the workplace and in communities. The National System of Education, implemented in the early 1980s, spanned child to adult, part-time to full-time, and literacy to technical educational programs.

The number of secondary schools quickly quadrupled. The number of adult educational and vocational centres doubled. A national university established in 1962 was renamed Eduardo Mondlane University in 1976 to honour the first president of Frelimo. It offers courses through a range of faculties, centres, and schools. Mozambicans trained both abroad and within the national system have increasingly assumed faculty positions, thus diminishing the nation's dependence on foreign faculty. In 1975 the ratio of Mozambican to foreign faculty was about 1 to 30, but by the early 1990s Mozambicans outnumbered foreign faculty at all levels.

The shortage of teachers, teacher dissatisfaction with low salaries, and conditions in refugee camps all tended to undermine the significant educational gains.

Health and welfare. Frelimo expanded and altered the health-care system to meet the needs of the majority population through rural health-care clinics. It changed the emphasis of the health-care system from curative care designed principally to serve the settler population to preventive care aimed at the majority. Frelimo's nationalization of medical services led to an exodus of licensed physicians. Despite initial progress, drought, armed assaults, population displacement, and general economic collapse since the early 1980s have led to a steady decline in the ratio of medical personnel per capita. The single exception is midwives, whose numbers have more than doubled since then. The overall trend has been a sharp decline in health and welfare conditions. The nation's major health problems derive from population displacement and the constraints of a tropical environment. Malnutrition, tuberculosis, cholera, common childhood diseases, a variety of gastrointestinal problems, and tropical diseases such as malaria, leprosy, schistosomiasis, and sleeping sickness are endemic in many parts of the country. By the beginning of the 21st century AIDS was a major health threat.

Cultural life. Mozambique enjoys a great range of cultural and linguistic diversity. Islāmic culture, Swahili language, and matrilineal Bantu-speaking groups coexist in northern and central regions, reflecting prevailing patterns in neighbouring Tanzania and Malaŵi. The great variety of people of the Zambezi valley overlap culturally and linguistically with neighbouring Malaŵi, Zambia, and Zimbabwe, and patrilineal, cattle-keeping people who share a heritage with neighbouring Nguni-speaking groups in South Africa and Zimbabwe are common in the south. Amid the variety of languages, social relationships, artistic traditions, clothing, and ornamentation patterns is a common theme of dynamic and creative cultural expression in song, oral poetry, dance, and performance.

Cultural diversity

Although material and performance arts are deeply embedded in daily religious and social expressions, some regional traditions are well known throughout the nation and beyond. The haunting paintings of Malangatana Valente Ngwenya, commonly known as Malangatana, have captured an international audience. Malangatana and the muralist Mankew Valente Muhumana have inspired the formation of artist cooperatives, particularly around Maputo. The carved wooden sculpture and masks of the Makonde people of northern Mozambique and Tanzania and the complex Chopi orchestral performances, or *midogo,* are among the best-known artistic traditions. Popular music includes the work of Alexandre Langa, Xidimingwana, and the Nampula group Eyuphuro.

Soccer is the nation's favourite sporting activity. Mozambique's soccer team competes with other African nations

and within the Portuguese-speaking Sporting League, which also includes Angola, Portugal, and Brazil.

Publishing

From the first decade of the 20th century, African writers and journalists published their own newspaper in the capital city. Despite problems of colonial censorship, the paper provided a forum for African intellectuals and writers throughout the century. Mozambicans studying abroad during the 1950s contributed to the literary and artistic flowering best known by the French term *négritude.* Writers used the colonial language to convey the experience of the colonized and to reconfirm the validity of African cultural expression. Some of Frelimo's leading figures, such as Marcelino dos Santos and Sérgio Vieira, wrote poetry and encouraged poetic expression as a tool of resistance. Mozambique's best-known writers in Portuguese include Luís Bernardo Honwana, José Craveirinha, and Orlando Mendes. The linguist and short-story author Bento Sitoe writes in Tsonga. The Association of Mozambican Writers sponsors seminars and public readings and publishes for the national market. The publishing group at Eduardo Mondlane University and the Historical Archive publish scholarly journals, monographs, edited collections, archival guides, and collections of documents.

With independence Mozambican cultural institutions underwent a fundamental transformation; although some institutions remain closed, most have eventually reopened in a different form. The Historical Archive of Mozambique, the Museum of the Revolution, the National Money Museum, the Geological Museum, and the Natural History Museum are the principal museums, archives, and libraries.

Despite Frelimo's emphasis on pride in African cultural heritage, its ideology of scientific materialism clashed sharply with important components of that heritage until the late 1980s. Spirituality, herbal and faith healing, rites of passage, direct criticism of leadership through poetic performance, and lineage authority over women all contradicted government efforts to reorder society along socialist lines and to define national culture through government-controlled newspapers, radio, publishing, and television. The government owns and controls most of the printed media, including *Notícias,* the daily national paper; *Tempo,* the weekly magazine; and *Domingo,* the Sunday paper. The Mozambique Information Agency is the country's official national and international news agency. Locally, Frelimo's Office of Mass Communications in Maputo has developed radio messages, murals, and a cartoon figure called Xiconhoca ("Chico the Snake," named after a despised agent of the dreaded colonial secret police), the embodiment of negative social attitudes, to convey its social message and to encourage communication with and among the nonliterate population.

For statistical data on the land and people of Mozambique, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

During the colonial era Mozambique's history was written as though it began with the arrival of the Portuguese, but the people of this region had developed complex communities based on agriculture, cattle keeping, mining, crafts, and trade long before the first small groups of Portuguese settlers arrived in the 16th century. Recent archaeological and historical research has begun to reintegrate Mozambique's past into that of eastern, central, and southern Africa.

Early settlement. From at least the 3rd century AD, Iron Age people who practiced agriculture and kept both cattle and small livestock moved into Mozambique as part of the migration of Bantu speakers from west-central Africa toward the south and east. Bantu speakers of the western African forest zone had mastered iron technology and combined the cultivation of some grains with knowledge of root and tree crops to sustain population growth to the point of expansion. In a slow but fairly steady process, one branch of Bantu speakers moved east toward the Indian Ocean and then south along the coast, and another moved more directly south-southeast into the Zimbabwe plateau and highlands of western Mozambique.

The characteristic social unit was an extended patrilineal household headed by an elder male and including his wives, their unmarried children, adult sons, and the sons' families. Although both social and labour organization varied throughout the area, women usually tended to child care, food preparation, cultivation, and gathering of food crops, whereas men were involved in cattle keeping, hunting, toolmaking, and a range of crafts.

Toward the end of the 1st millennium in south-central Mozambique, groups of households called *nyika* had emerged as social units under the authority of a chief and chiefly household. In the 10th century a settlement known as Mapungubwe, which incorporated many *nyika,* developed in the upper reaches of the Limpopo River. It was the earliest known of the settlements featuring stone enclosures, or *zimbabwes.* Mapungubwe revealed marked social differentiation as early as the 10th to the 11th centuries.

Mapungubwe

The rise of the *zimbabwe* civilizations. The groups of the Zimbabwe plateau expanded their herds and moved between the plateau and the surrounding Mozambican lowlands in pursuit of seasonal pasturage, although the tsetse fly was present in the region. On the Zimbabwe plateau the stone-enclosure settlements reached their peak prosperity. The region's economy was rooted in agriculture and cattle keeping, but its social and political organization became more complex with the development of local industries and trade, specifically the mining of gold, copper, and iron ore and the development of salt pans, tool forges, and potting industries. The civilization of Great Zimbabwe, which dominated the region politically from the mid-13th to the mid-15th century, controlled mining and trade.

The ancient *zimbabwe* settlement at Manekweni, about 30 miles from the Indian Ocean in southern Mozambique, replicated the social and settlement patterns of the highland interior in miniature. Manekweni was a centre for agriculture, cattle keeping, and the gold trade from about the 12th to the 18th century.

From about the 10th to the 18th century, Great Zimbabwe and the area of central Africa around Lake Kisale (in the present-day Katanga province of Zaire) were the region's centres of production and intra-African trade. From at least the 1st millennium, however, people of this region traded with various non-Africans. The earliest and most important external trade link for Mozambicans was with Middle Eastern and Asian people who traded beads and cloth for gold across the Red Sea and Indian Ocean. People of the present-day territory of Mozambique participated in all three trades: the gold and cattle trade of the Zimbabwe highlands; the pottery, salt, and fish trade of the Lake Kisale region; and the bead, cloth, and gold trade between the Zimbabwe plateau and the East African coast.

Growth of external trade

By the 14th century, Afro-Arab, or Swahili, trade cities flourished along the coast from Somalia in the north to Kilwa in southern Tanzania. Smaller Swahili sultanates developed along the northern coast of Mozambique as far south as Angoche, and Arabs traded with Shona, Tsonga, and Yao traders in outposts from Mozambique Island south to Sofala. Early 14th-century pottery from Kilwa and Manda, Swahili city-states hundreds of miles north on the Tanzanian coast, has been found at Chibuene, near Manekweni, confirming that the East African coastal trade extended far into southern Mozambique quite early. By the 16th century, intra-regional trade in raw materials and long-distance trade in gold, copper, ivory, and some slaves sustained a series of markets throughout the region.

Arrival of the Portuguese. The voyage of Vasco da Gama around the Cape of Good Hope into the Indian Ocean in 1498 marked the European entry into trade, politics, and society in the Indian Ocean world. During the early 16th century, the Portuguese established their dominance at Mozambique Island and Sofala. By the 1530s small groups of Portuguese had pushed their way into the interior, setting up garrisons and trading posts at Sena and Tete on the Zambezi River, in an effort to gain exclusive control over the gold trade. The Portuguese attempted to legitimate and consolidate their trade and settlement positions along the Zambezi through the creation of *prazos* (land grants) tied to European occupation.

The *prazos* developed as Afro-Portuguese or Afro-Indian centres defended by large slave armies. Slave trading and Nguni attacks in the first half of the 19th century undermined existing *prazos*, but several heavily fortified *prazos* survived and strongly resisted Portuguese domination until the last quarter of the 19th century.

Between the 16th and 19th centuries control over trade, tribute, and production in central and northern Mozambique clustered around a series of Shona polities south of the Zambezi and Marave polities north of the river. *Prazo* holders and Arab and Portuguese traders all tried to advance their own positions in what remained a fairly fluid situation. The Portuguese were able to wrest much of the coastal trade from Arabs between 1500 and 1700, but with the Arab seizure of Portugal's key foothold at Fort Jesus on Mombasa (now in Kenya) in 1698 the pendulum began to swing in the other direction. During the 18th and 19th centuries the Mazrū'ī and Omani Arabs reclaimed much of the Indian Ocean trade, forcing the Portuguese to retreat south. During the 19th century other European powers, particularly the British and the French, became increasingly involved in the trade and politics of the region.

Effects of the slave trade. By the 18th century, slaves had become an increasingly important part of Mozambique's overall export trade from the East African coast. Yao traders developed slave networks from the Marave area around the tip of Lake Nyasa to Kilwa and Mozambique Island. *Prazo* traders along the Zambezi sold gold and slaves from Zumbo, Tete, and Manica to Portuguese merchants at Quelimane, and Tsonga ivory traders developed routes from the Transvaal and Zimbabwe plateau to coastal entrepôts at Inhambane and Lourenço Marques (now Maputo). During the 19th century, Mozambicans were sold as slaves in the Portuguese and Brazilian South Atlantic trade, the Arab trade from the Swahili coast, the French trade to the sugar-producing islands of the Indian Ocean, and to Madagascar. Although the trade in slaves declined as a result of the mid-19th-century slave-trade agreements between Portugal and Britain, clandestine trade, particularly from central and northern Mozambique continued into the 20th century.

Formation of the Gaza state

During the first 30 years of the 19th century, the proliferation of military and raiding groups from the conflicts in the northern Nguni heartland southwest of Mozambique had an important impact on southern and west-central Mozambique. Several military groups, offshoots of the emerging Zulu state, invaded Mozambican territory, seizing cattle, hostages, and food as they went. The waves of armed groups disrupted both trade and day-to-day production throughout the area. Two groups, one under Zwangendaba and the other under Soshangane, swept through Mozambique. Zwangendaba's group continued north across the Zambezi, settling to the west of contemporary Mozambique, but Soshangane's group crossed the Limpopo into southern Mozambique, where it eventually consolidated itself into the Gaza state. In the 1860s, a succession struggle between the sons of Soshangane caused enormous suffering in the region and weakened the Gaza state.

Consolidation of Portuguese control. The Gaza pattern of raiding groups that refused to pay tribute led some of its neighbours to view alliance with the Portuguese as an attractive alternative to periodic plunder. The Portuguese exploited such tensions. In the 1890s Portugal mounted a coalition of Portuguese troops and African armies against the Gaza state. With the defeat and deportation of Gaza leadership between 1895 and 1897, southern Mozambique passed into Portuguese control.

Trade in ivory, gold, slaves, rubber, oilseeds, and a broad range of European trade goods continued throughout the 19th century. However, European economic interest and influence in the region began to change rapidly in response to developments in both Africa and Europe. By the 1860s people of southern Mozambique began selling their labour at the sugar plantations and ports of South Africa instead of hunting elephants or gathering oilseeds. With the major South African mineral strikes in the 1860s (diamonds at Kimberley) and 1880s (gold at Witwatersrand), European interest in African labour power superseded European interest in African-produced trade goods. With that shift came the European determination to wrest greater control over land and labour from African leadership. The combined struggle for access to mineral-bearing soils and the labour force to work them fueled the so-called "Scramble for Africa" in southern Africa.

Portugal put forth its claim to the entire swath of Africa from contemporary Mozambique to Angola. The Germans, whose territory bordered Mozambique to the north, accepted the Portuguese claims, thus establishing Mozambique's northern boundary. British claims to the region contradicted Portugal's, leading to prolonged negotiations. The Portuguese crown was heavily in debt to British financiers, and the small nation was no match for Britain's military. In 1891 Portugal was forced to bow to Britain's definition of Mozambique's western and southern boundaries.

Contested boundaries

Mozambique's boundaries may have been set through European litigation and diplomacy, but they were fiercely contested by Mozambicans. In the late 19th century there was little to suggest that Mozambique was a united nation. By 1890 the Gaza state was the most influential polity in southern Mozambique. Other important groups included the Barue of central Mozambique, the Afro-Portuguese of the Zambezi *prazos* and the *aringa* of Maganja da Costa, the Yao of Mataka, the Makua chieftaincies throughout the north, and the northern coastal sheikhdoms of Angoche. The Portuguese controlled trade and collected tribute in coastal enclaves from Ibo in the north to Lourenço Marques in the south, but their ability to control events outside these areas was quite limited. From the 1880s to 1917 the Portuguese mounted dozens of military campaigns to extend and confirm their control.

Portugal had little hope of developing the entire region on its own and so turned to the then familiar colonial strategy of leasing great tracts of the area to private companies. Chartered companies were granted the privilege of exploiting the lands and peoples of specific areas in exchange for the obligation to develop agriculture, communications, social services, and trade. The Mozambique Company, the Niassa Company, and the Zambezia Company were established in the 1890s. The limited development and investment in infrastructure that occurred related directly to company interests and were usually undertaken at African expense. Labour conscription for road and railway building, plantation field labour, and the provision of contract labour for the mines and plantations of neighbouring areas superseded slavery in much of the area and promoted social disruption. Sugar, copra, and sisal plantations depended largely on conscript labour, and railways linking Beira with the British South Africa Company territory and British Nyasaland to the west and northwest were built at a high cost to African labour.

The Portuguese government eventually terminated the charters of the major concession companies, bringing all Mozambique under direct Portuguese rule. The most important colonial abuses—forced labour, forced crop cultivation, high taxes, low wages, and alienation of the most promising lands—were as prevalent in areas under direct Portuguese rule as they had been during the companies' rule. Between the 1890s and the 1930s, Portuguese rule in Mozambique was characterized by the exploitation of people and resources by private parties, whether foreign company shareholders or colonial bureaucrats and settlers.

Mozambique under the "New State" regime. The coup of 1926 ushered in a regime in Portugal that came to be known as the "New State" (Estado Novo). Although most of the former abuses in Mozambique continued and in some cases were intensified, the New State consolidated the profit into fewer hands. Protectionist policies instituted in response to the economic crisis of the 1930s gave Portuguese investors and settlers an edge over foreign capital. The administrative and educational systems were unified and developed more coherently under the New State, but they were still principally directed toward settlers. The New State promoted conditions in Mozambique that would allow Portugal and the Portuguese to accumulate capital.

Protectionist policies

Colonial investment patterns began to change in the early 1950s. Portugal set forth a series of development plans designed to extend and upgrade Mozambique's transportation and communications infrastructure. Portuguese who had exploited monopolies and incentives to accumulate capital were encouraged to invest in, expand, and diversify their undertakings. The generally favourable prices for tropical commodities in the postwar era fueled the trend, and the colonial economy expanded quite vigorously. The New State retained tight control over African economic and physical mobility. In the 1950s and '60s thousands of Portuguese settlers arrived to take advantage of employment and business opportunities denied to Mozambicans. By closing off opportunities for the most upwardly mobile, educated, and motivated Africans and maintaining strong discipline and pressure on African farmers and day labourers, Portuguese colonial authorities antagonized Mozambicans across the class spectrum.

By the late 1950s, African leadership emerged from a cross section of the population to channel discontent against the colonial power. The formal political challenge developed among Mozambican workers and students living outside Mozambique because the New State met overt political dissent of any character with imprisonment, death, or deportation. In 1962 Mozambican representatives from exiled political groups met in Tanganyika and forged the Mozambique Liberation Front, or Frelimo (Frente de Libertação de Moçambique).

Formation of Frelimo

Frelimo's strategy was not immediately clear, but, after serious dissent within the organization, the ascendant leadership committed itself to an armed challenge. In September 1964, Frelimo's guerrilla forces, which had been trained and armed by African and Eastern-bloc supporters, attacked targets in northern Mozambique, and the war for independence was launched. Portugal, faced with similar challenges in all its African territories, responded by fielding an enormous military effort. The Portuguese and Frelimo each experienced important victories and setbacks during the ensuing decade of struggle. Although they inflicted some important damage on Frelimo forces, the Portuguese remained frustrated and offensively ineffective against Frelimo's small-scale guerrilla engagements. By 1974 Frelimo forces could move about fairly freely in most of the north and had infiltrated into central Mozambique. The cities, most of the south, and the coastal areas as far north as Nacala remained in Portuguese hands.

On April 25, 1974, the Armed Forces Movement in Portugal staged a coup d'état, which was welcomed by Portuguese discontented with the New State regime, its African wars, and its backward-looking ideology. Frelimo took advantage of its military position to insist on a ceasefire, which confirmed its right to assume power in an independent Mozambique. A quickly aborted countercoup attempt in the capital in September and some rioting in October were the only overt challenges to Frelimo's authority. Within a year of the Portuguese coup, most of the settler population had left Mozambique. On June 25, 1975, Mozambique became independent under Frelimo's single-party rule. (J.M.Pe.)

Independence. *Mozambique as a one-party state.* Mozambicans widely supported Frelimo's decisions to implement economic sanctions against Rhodesia (now Zimbabwe) and to allow guerrilla forces opposed to the Rhodesian government to develop bases in Mozambique, but Mozambique suffered major losses of revenue and lives and the destruction of key infrastructure. Frelimo's support for the African National Congress (ANC) brought similar economic and military retribution from South Africa.

Frelimo had mixed success with its social and economic policies during its first decade of rule. While forced cultivation, forced labour, and ethnic discrimination were ended, the party's commitment to communal and state-run agriculture antagonized many farmers, who had hoped to see land returned to their families. By 1985 Frelimo recognized the failure of its agricultural policy and, under pressure from international creditors, began to favour the family agricultural sector over state-run markets.

The government's extensive investment in education, health care, and services was initially highly successful. Within a decade of independence, however, these gains had been totally undermined by the actions of the Mozambican National Resistance (Resistência Nacional Moçambicana; Renamo), an insurgency group trained, supplied, and supported largely by Rhodesia and South Africa. Renamo began a campaign of terror against the rural population and economic sabotage shortly after independence. A 1984 agreement between Mozambique and South Africa to end support for guerrilla groups did little to curb Renamo's activity, and Frelimo continued its attempts to end the conflict through negotiation.

Renamo's campaign

Mozambique since 1992. The leaders of Frelimo and Renamo finally accepted a peace accord in October 1992, whereby Frelimo's leadership agreed to change the constitution and to open the political process to competing parties in exchange for Renamo's promise to end the war. Issues between the two groups were not totally resolved until 1994, and multiparty elections in that year were the culmination of compromises on both sides. Frelimo ended its one-party rule and Mozambique's identity as a socialist nation. Renamo changed its image as an international pariah known for burning schools and health clinics and became a legitimate political party. Renamo's leader, Afonso Dhlakama, was accepted as a legitimate presidential candidate. Although Frelimo and Renamo were the main contenders in the elections, many other smaller parties also emerged. There were a dozen presidential candidates and a similar number of parties fielding candidates for the National Assembly. The elections were considered free and fair by international observers, with Frelimo president Joaquim Chissano garnering the majority of the votes. Renamo, however, was strongly represented in the national government and the National Assembly.

The new government faced the lingering effects of the war, notably the presence of up to two million land mines in the countryside. The effort to demobilize both Frelimo and Renamo forces and form a new unified military also met with delays and difficulties. In the mid-1990s soldiers waiting in demobilization camps for weeks without food, money, or prospects for work staged scattered violent uprisings. A major concern of the government at the end of the 1990s was how to reform land tenure. (J.M.Pe./Ed.)

For later developments in the history of Mozambique, see the BRITANNICA BOOK OF THE YEAR.

Namibia

The Republic of Namibia (formerly called South West Africa) is a vast (318,580 square miles [825,118 square kilometres]), nearly empty land bordered by Angola to the north, Zambia to the northeast, Botswana to the east, South Africa to the southeast and south, and the Atlantic Ocean to the west. It ranges from arid in the north to desert on the coast and in the east. The landscape is spectacular, but the desert, mountains, canyons, and savannas are perhaps better to see than to occupy.

The only permanent rivers are the Kunene (Cunene), the Okavango (Cubango), the Mashi (Kwando), and the Zambezi on the northern border and the Orange on the southern. Only the northern frontier—and not all of it—is readily passable. The coastal Namib desert, the treacherous reefs and shoals of the coast (half aptly named the "Skeleton Coast"), the near deserts along the Orange River, and the dry Kalahari region to the east explain the late conquest of Namibia and form a geographic frame around the country. Roughly rectangular (600 by 300 to 450 miles [965 by 480 to 725 kilometres]), Namibia has a long, narrow eastern extension (the Caprivi Strip). Namibia became independent on March 21, 1990, under a democratic multiparty constitution. The capital is Windhoek.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Namibia is divided from west to east into three main topographic zones: the coastal Namib desert, the Central Plateau, and the Kalahari. The Namib is partly rocky and partly (in the central stretch) dunes. While having complex flora and fauna, it is a fragile and sparsely covered environment unsuitable for pastoral or agricultural activities. Diamonds (probably washed down

from the Basotho highlands by the Orange River) and uranium are found at Oranjemund in the south and Arandis in the centre. The Namib, 50 to 80 miles wide over most of its length, is constricted in the north where the Kaokoveld, the western mountain scarp of the Central Plateau, abuts on the sea.

The Central Plateau, which varies in altitude from 3,200 to 6,500 feet (975 to 1,980 metres), is the core of the agricultural life of Namibia. In the north it abuts on the Kunene and Okavango river valleys and in the south on the Orange. Largely savanna and scrub, it is somewhat more wooded in parts of the north and is broken throughout by hills, mountains, ravines (including the massive Fish River Canyon), and salt pans (notably the Etosha Pan). Mount Brand (8,445 feet [2,574 metres]), Namibia's highest peak, is located along the plateau's western escarpment.

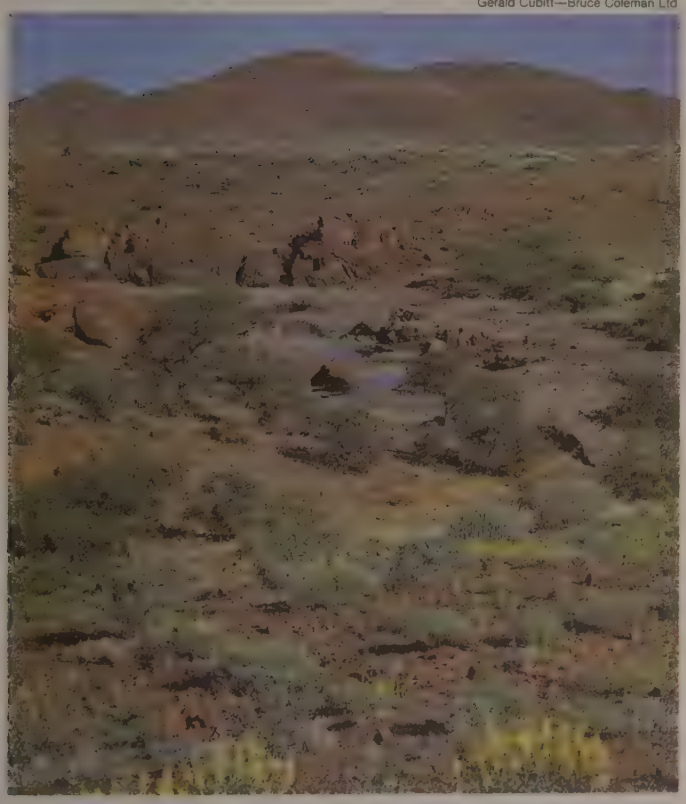
In the east, Namibia slopes gradually downward, and the savanna merges into the Kalahari. In the north, hardpan and rock beneath the sand, in addition to more abundant river water and rainfall, make both herding and cultivation possible.

Intermittent streams

Drainage and soils. As noted, only the border rivers are permanent. The Swakop and Kuiseb rivers rise on the plateau, descend the western escarpment, and die out in the Namib (except in rare flood years, when they reach the sea at Swakopmund and Walvis Bay, respectively). The Fish (Vis) River rises in the Central Plateau and (seasonally) flows south to the Orange. Various lesser rivers rise on the plateau and die out downstream in the Namib or Kalahari desert.

Namibia's soils range from barren sand and rock to low-quality sand-dominated to relatively fertile soils. The best soils are in the north, in the Otavi Mountains, in parts of the central and southern portions of the plateau, and in the Caprivi Strip. Water—not soil fertility—is the primary constraint on agriculture. Both in the densely populated Ovambo region in the north and in the commercial farming areas, overuse of land has reduced tree and bush cover, compacted soils, led to serious erosion, and lowered the water table by as much as 100 feet in the 20th century.

Gerald Cubitt—Bruce Coleman Ltd



The intermittent Uniab River in its rainy-season course downstream from its headwaters on the Central Plateau (near Palmwag) through the rocky terrain and scrub bush of the Namib (heading, if rain persists, to the Atlantic Ocean, near Torra Bay, Namibia).

Climate. Namibia is located on the southern margin of the tropics and has distinct seasons. The coast is cooled by the Benguela Current (which carries with it the country's rich and recovering fish stocks) and averages less than 2 inches (50 millimetres) of rainfall annually. The Central Plateau and the Kalahari have wide diurnal temperature ranges, more than 50° F (30° C) on summer days and less than 20° F (10° C) in winter. In Windhoek, on the plateau, the average temperature for December is 75° F (24° C), and the average maximum 88° F (31° C). In July these averages are 55° F (13° C) and 68° F (20° C), respectively. Humidity is normally low, and rainfall increases from about 10 inches (250 millimetres) on the southern and western parts of the plateau to about 20 inches in the north-central part and more than 24 inches on the Caprivi Strip and Otavi Mountains. However, rainfall is highly variable, and multiyear droughts are common. In the north and adjacent to mountains, groundwater is as important as—but only slightly less variable than—rainfall. Kalahari rainfall—in its Namibian portion—is not radically different from that of the plateau, but, except in the northern Karstveld and isolated artesian areas, groundwater is less available.

Plant and animal life. Both the Namib and Kalahari deserts are characterized by exotic, fragile desert plants. The mountains are sparsely wooded, and the plateau is predominantly scrub bush and grass. Trees are much more frequent in the north. Varieties of aloe are common throughout the plateau and the less sandy portions of the Kalahari.

Namibia is richly endowed with game, albeit poaching has seriously diminished it in parts of the north. Throughout the ranching zone, game (notably antelope and giraffes) coexists with cattle and sheep. The Etosha Pan in the north is a major game area and tourist attraction.

Settlement patterns. Less than 1 percent of the country is estimated to be arable, though almost two-thirds is suitable for pastoralism. Wasteland (mountain and desert) and bush or wooded savanna, plus a small forest zone, constitute the remainder.

Population distribution

About half of the entire population live in the far north, roughly 15 percent in the commercial ranching areas north and south of Windhoek, 10 percent in central and southern ex-black homelands, more than 10 percent in Greater Windhoek, and the remainder in coastal towns and inland mining towns. More than one-fourth of the total population live in urban areas. Namibia's population is young—about half are 16 years of age or younger—and is growing at a relatively modest rate compared with those of other African countries.

The people. *Ethnic and linguistic composition.* About 85 percent of Namibians are black, 5 percent of European ancestry, and 10 percent, in South African terminology, Coloured (Cape Coloured, Nama, and Rehobother). Of the black majority, about two-thirds are Ovambo, with the Kavango, the Herero, the Damara, and the Caprivian peoples following in population size. Other ethnic groups have much smaller populations. Afrikaners and Germans constitute two-thirds and one-fifth of the European population, respectively. Most ethnic Europeans are Namibian citizens, though some have retained South African citizenship.

English is the national language, though it is the home language of only about 3 percent of the population. Ovambo languages are spoken by more than 80 percent of the population, followed by Nama-Damara with about 6 percent. Kavango and Caprivian languages and Herero, as well as Afrikaans, constitute about 4 percent of home languages. Many Namibians speak two or more indigenous languages and at least a little of two of the three European languages (English, Afrikaans, German) in common use.

Religion. Some 80 to 90 percent of the population at least formally adheres to a Christian confession. The largest denominations are two Lutheran churches, which together encompass about one-half the total population. Roman Catholics comprise another one-fifth of the population, while the Dutch Reformed and Anglican denom-

inations each make up small percentages of the total. There are still smaller groups in the African Methodist Episcopal, Methodist, and Presbyterian churches, and a small percentage practices traditional religions.

Demographics. Namibia's population density is among the world's lowest, and about two-thirds of the people live in rural areas. The country's annual population-growth rate is slightly higher than the regional average. Its birth rate is much higher than the world average, and its death rate is almost twice that of the world average. The average life expectancy of about 45 years is down from 56 in the early 1990s, a result of the swift spread of HIV/AIDS. This circumstance has concurrently produced a rapidly increasing number of orphans (more than 82,000 were registered in 2004) who have lost parents to the disease. In part because the disease is easily transmittable to nursing babies, infant mortality remains a serious problem; the overall rate is higher than the average for countries in southern Africa but below that of most countries in sub-Saharan Africa. Two-fifths of the population of Namibia is under 15 years old.

Like almost all other human welfare indicators, the black-white disparity in demographics is very high—a legacy, in large part, of the South African occupation regime's practice of apartheid. Many exiles have returned to the country since independence, whereas some 10,000 Europeans and almost as many black South African hired troops and auxiliaries have departed.

The economy. Nominally Namibia is a lower-middle-income economy with a gross domestic product (GDP) per capita that is significantly above average for countries in sub-Saharan Africa. But that summary is misleading. Only one-quarter of all Namibians and only one-sixth of black Namibians have adequate incomes; up to two-thirds live in abject poverty with limited access to public services. More than a third of the population is unemployed, and overall economic growth is difficult because of a shrinking productive sector, lack of capital stock, and extreme fluctuations in the world market for base metals and uranium oxide.

Problems in economic growth

Agriculture and fishing. Commercial farming (undertaken predominantly by white settlers) was concentrated on the raising of Karakul sheep (sheepskins) and cattle (beef) for export. Crop raising (mostly beans, potatoes, millet, and maize [corn]) is a distinctly secondary activity on commercial farms, but it is almost equal with livestock production on small family farms in the north. Many of the latter operate at below subsistence level and are headed by women. Rural development efforts are aimed at small farmers. From the start of independence the Namibian government has attempted to redistribute enough of the arable land to help eliminate the extremes of rural poverty. White farmers, by most accounts, held well over half the country's arable land. In 1995 the government began what was called a "willing seller–willing buyer" program. If, for example, a farmer held many farms and would be willing to sell off one or some of these, the government had right of first refusal. This program also allowed the government to purchase unused land. But progress in land redistribution was slow, and in 2004 the government announced that it would begin expropriating land, making every effort to give the owners fair compensation.

Commercial fishing (pilchards [sardines], anchovy, hake, horse mackerel, oysters, and mussels) and fish processing (an important source of jobs) are a rapidly growing sector of the Namibian economy.

Mining

Industry. Mining is central to the economy, though the government was attempting to diversify. Extractive industries account for more than one-tenth of the GDP. Gem diamonds, uranium oxide, and base metals dominate mining; however, gold and natural gas are increasingly significant, and oil production (offshore and in the Etosha basin) is potentially so. Namibia supplies nearly one-third of the world diamond output, but the value of this contribution varies with world prices. Uranium and copper are also produced in significant amounts. Other important minerals include fluorite, lead, zinc, cadmium, pyrite, tin, and silver.

Manufacturing produces about one-tenth of the GDP. It is dominated by cut gems, fur products, processed foods, textiles, carved wood products, and refined metals.

The beauty and diversity of the Namibian landscape—especially on the coast, at Etosha, and in Fish River Canyon—combined with the country's relative stability, made tourism an expanding industry in the 1990s, and by the turn of the 21st century it was the fastest-growing sector of the economy.

Finance and trade. Namibia has several commercial banks, but three of them—First National Bank of Namibia, Standard Bank Namibia, and Bank Windhoek—account for most banking business. Reorganization of land, housing, and development banks was begun after independence. The Central Bank of Namibia launched an independent currency, the Namibian dollar, to replace the South African rand in the mid-1990s.

South Africa is by far the largest source of Namibia's imports, supplying food, beverages, and tobacco; machinery; transport equipment; and base and fabricated metals. The United Kingdom, South Africa, Spain, and France are the major export destinations for diamonds and other minerals, processed fish, live animals and animal products, and beverages and other food products.

Transportation. Transportation is dominated by TransNamib, a public-sector rail, road, and airline operator. Transport infrastructure is reasonably good, with main routes through the Caprivi Strip (and thence to Zambia and Zimbabwe) and to Botswana being upgraded. Air Namibia flies to national and regional destinations and to Europe. There is an international airport at Windhoek. A handful of large road-transport companies compete with larger numbers of small haulers.

Administration and social conditions. *Government.* Namibia is a multiparty democracy with a president as head of state. Legislative power is vested in the National Assembly, which consists of 72 members directly elected to five-year terms under universal adult suffrage; the president may appoint up to six additional (nonvoting) members. The work of the assembly is subject to review by the National Council, which is made up of members elected by regional councils.

The constitution

Namibia's constitution, which took effect at independence, is highly rights-conscious and aimed at achieving a durable separation of powers. Arguably, the results have a complexity and rigidity (many clauses literally cannot be amended) that may hamper governance. They represent a compromise in which individual (not group) rights and affirmative action are included in return for a series of safeguards against majority legislative or executive abuse of power and entrenchment of property rights.

The country's armed forces have been formed largely from the military wing of the South West Africa People's Organization (SWAPO), the People's Liberation Army of Namibia. The police forces, which are about the same size as the army, remain primarily a holdover institution from preindependence times, which has led both to a certain lack of confidence in them and to a popular reaction of caution—neither attitude being helpful in controlling Namibia's postwar upsurge in crime.

Namibia was admitted to the United Nations in 1990. It joined several regional organizations, including the Organization of African Unity (now the African Union), the Southern African Development Coordination Conference (now the Southern African Development Community), and the South African Customs Union. The latter membership, which allows the open and duty-free exchange of goods between South Africa and Namibia, is an indication of the pragmatic and largely noncontentious relations the two countries have maintained. Namibia also has membership in a number of global organizations, such as the African, Caribbean, and Pacific (ACP) group of developing nations; the World Trade Organization; the International Monetary Fund; and the Commonwealth.

Education. The government's policy of expanding education to offer universal primary and lower secondary instruction faces severe problems. Most teachers, for example, are far from qualified (not least in the language of instruction, English). Yet more than four-fifths of all

children between the ages of 7 and 18—a figure far higher than in most African countries—are enrolled in school. The student-teacher ratio is also comparatively low, even for southern Africa, which tends to have a lower ratio than most other regions of the continent.

With more than 70 percent of its adult population literate, Namibia has one of the highest rates of literacy in sub-Saharan Africa. Various informal adult education programs have been implemented to combat the remaining illiteracy. Higher education is provided by four teacher-training colleges and a university.

Health and welfare. Most Namibians are poor—more than half are absolutely poor—and nutritional standards are low. As much as one-tenth of the population suffers severe, chronic malnutrition. Formal wage employment engages only a small percentage of the work force, and unemployment is high. Average white incomes are several times higher than formal, waged black incomes, but the non-wage-generated incomes of most blacks fall far short even of this low level.

Namibia has one of the best health-care systems in Africa, as measured by both its population-to-doctor and its population-to-hospital-bed ratios. Emphasis is placed on primary and preventative health services, and the country's system of regional hospitals and mobile clinics has attempted to raise the level of services available in rural locations.

Social policy

Women and children have received special attention in social policy, being the most disadvantaged groups. In the case of women, ending legal and social discrimination and improving access to education, land, and employment are stated goals toward which some action has begun. A detailed government paper laid out ways and means to meet the child health, education, nutrition, and other goals for the year 2000 adopted by the 1990 World Summit on Children.

Cultural life. Namibian cultures are diverse. Just as the culture of the Afrikaners differs significantly from that of the German-speaking community and as both of those cultures differ from that of the more varied technical-assistance community, so do African and Creole cultures differ. The Rehobothers closely resemble the rural Afrikaner culture of the mid-20th century, while the Nama have more in common with the other pastoral black communities, and the "Cape Coloured" have a distinct urban culture with both black and European elements. The northern black cultures—while distinctive as to language and forms of music and dance—formed out of a mixed farming context unlike that of the Damara and Herero. The San are a tragic case. Their culture was ruined by ranch serfdom and wartime exploitation as trackers, and efforts to rebuild from the fragments have been limited by lack of knowledge, resources, and space as well as by the paternalism of many of their self-appointed "guardians."

With the exception of the San, Namibian cultures appear to be alive and evolving, not least in the urban areas. However, rising unemployment may lead to the breakdown of neighbourhood and other social groupings and to the anomie and lawlessness that characterize the townships of many southern African cities, notably in both Zambia and South Africa. The black cultures are not well supported by formal institutions or the government, owing both to doubts as to what would enable rather than smother their development and to a lack of fiscal resources.

A number of holidays and festivals are observed, most of which are religious or historic in significance, albeit not necessarily of specific current content. Sports are popular among both spectators and participants. A wide variety of sports are followed by the white communities, but the black communities concentrate on soccer.

Radio and television broadcasting services are government-owned, as is one daily newspaper. All appear to have substantial intellectual and programmatic freedom. A fluctuating band of party, semi-party, and (in one case) independent newspapers exist and are not subject to censorship, but the survival of most is in doubt for economic reasons. They are supplemented by an array of religious, trade union, and other specialized papers that also have complete freedom of expression.

For statistical data on the land and people of Namibia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The history of Namibia is not well chronicled. Its isolated geographic position limited contact with the outside world until the 19th century. Explorer, missionary, trader, conqueror, and settler sources are neither comprehensive, notable for accuracy, nor unbiased. Professional historiography is a post-1960 development in the country, and the political events of the years since then have coloured most of the written history.

Independence before the conquest. The earliest Namibians were San (often pejoratively called Bushmen), nomadic peoples with a survival-oriented culture based on hunting and gathering. Their clans were small and rarely federated, and their military technology was so weak that even before the arrival of the Europeans they had been pushed back to the desert margins.

First conquerors

The first conquerors in southern Namibia were the Nama (of a people often—though not in Namibia—pejoratively called Hottentots). They had a larger clan system, with interclan alliances, and a pastoral economy. Closely linked (usually in a dependent role) were the Damara, a people from central Africa whose culture combined pastoralism, hunting, and copper smelting. In northeastern and central Namibia the Herero (a pastoral people from central Africa) built up interlocked clan systems eventually headed by a paramount chief. The unity of the Herero nation, however, was always subject to splintering. In the north the Ovambo people developed several kingdoms on both sides of the Kunene River. They were mixed farmers (largely because of a more hospitable environment for crops) and also smelted and worked copper. To the east the related Kavango peoples had a somewhat similar but weaker state system. On the margins of Namibia—*i.e.*, the Caprivi Strip in the far east and on the margins of the Kalahari—the local peoples and groupings were spillovers from southern Zambia (Barotse) and Botswana (Tswana).

Until the 1860s, European contact and penetration were slight. Diogo Cão and Bartolomeu Dias touched on the Namibian coast in 1486 and 1488 respectively, en route to and returning from the Cape of Good Hope, but there was virtually no contact until the 1670s. Afrikaner explorers after 1670 and Afrikaner traders and settlers about 1790 came to Namibia and eventually reached the southern boundaries of the Ovambo kingdoms, notably at the Etosha Pan. They—together with German missionaries, explorers of varied nationality, British traders, and Norwegian whalers—did not play a dominant role before 1860. Instead, they created the first avenues for trade (ivory and later cattle) and introduced firearms.

The latter heightened the destructiveness of conflicts among the various clans and peoples. So did the arrival, after the first quarter of the 19th century, of the Oorlam-Nama from the Cape. Their military technology (which included horses, guns, and a small mobile commando organizational pattern) was modeled on that of the Afrikaners. They came to dominate the resident Nama (Red Nation) and Damara. In the middle of the 19th century, a kingdom ruled by the Oorlam but partly Herero and supported by the Red Nation and Damara was established near Windhoek by the Oorlam chief Jonker Afrikaner.

Central Namibia was then an area of conflict between the southward-moving Herero and the northward-migrating Nama. In 1870 a peace treaty was signed with the Germans on the border of Herero country. Meanwhile, largely as a result of war pressures, Maherero had emerged as the Herero paramount chief. At this time a South African Creole ("Coloured") community, the Rehoboth Basters, had immigrated to a territory south of Windhoek, where they served as a buffer between the Herero and the Germans. Like the Oorlam, they were Europeanized in military technology as well as civil society and state organization, which were copied from the Afrikaners.

The German conquest. In the 1870s, British annexation of Namibia appeared imminent. A treaty with the Herero and the raising of the British flag over Walvis Bay were

seen as forerunners of the northward expansion of the Cape Colony. However, London proved reluctant to take on added costs in an apparently valueless area, and the way was left open to German colonial annexation as South West Africa in the 1880s. The acquisitions, by exceedingly dubious "treaties" and more naked theft, did not go smoothly, despite the employment of so-called "divide and rule" tactics within and between peoples. The first major resistance—by the Herero in 1885—forced the Germans back to Walvis Bay until British troops were sent out.

By the turn of the century, German settlers had arrived, copper was minable, railway building from Swakopmund and Lüderitz was under way, and diamonds were soon discovered near Lüderitz. But from 1904 to 1907 a great war of resistance broke out, nearly expelling the Germans before it was quelled with extreme savagery by tactics including extermination, hangings, and forced detention in concentration camps.

War against the Herero

The first phase of the war was fought between the Germans and the Herero (with a single Ovambo battle at Fort Namutoni near the Etosha Pan). It reached a climax when General Lother van Trotha defeated the main Herero army at the Battle of Waterburg and, taking no prisoners, drove them into the Kalahari, where most died. By 1910 the loss of life by hanging, battle, or starvation and thirst—plus the escape of a few to the Bechuanaland protectorate—had reduced the Herero people by about 90 percent (80–85 percent dead, 5–10 percent in exile). The Nama resistance war came late because a key letter from Maherero's son and successor, Samuel Maherero, to the Oorlam chief Hendrik Witbooi that proposed joint action had been intercepted. The resistance was finally crushed in 1907, and Nama survivors were herded into concentration camps. War, starvation, and conditions in the camps claimed the lives of two-thirds of the Nama.

The Germans allocated about half of the usable—and apparently all of the best—ranchland (except that of the Rehoboth Basters) to settlers and restricted Africans to reserves. The Tsumeb copper and zinc mines opened in 1906, and diamond mining (more accurately, sand sifting) began near Lüderitz in 1908 and at the main fields at the mouth of the Orange River (Oranjemund) a few years later. Railways linked Lüderitz, Keetmanshoop, and Windhoek as well as Swakopmund, Windhoek, and Tsumeb.

German direct rule never extended to the north. The "red line"—now a quarantine boundary—delimited the Police Zone from the Ovambo and Kavango areas. In the latter, the near extinction of elephants, a rinderpest epidemic, and the rising consumption habits of the kings led to a migration of single male contract labourers to work in the mines and ranches and in construction. "Contract"—which was to provide the cheap labour for the colonial economy and later provided the national communication and solidarity links to build the liberation movements of 1960–90—had begun.

The Boer conquest. In 1914–15 South African troops invaded and captured South West Africa as part of the World War I conquest of the German colonies in Africa. Except for diamond mines, most property—including Tsumeb—found its way back into German hands. The rising De Beers colossus bought Oranjemund and the balance of the diamond-producing area to bolster its world domination; it was used as a market-balancing mine (that is, its production was varied to control the price of diamonds, and it was totally closed for more than two years in the 1930s), a role it played into the 1980s. Afrikaner settlers were encouraged to come to South West Africa for security reasons—to hold the inhabitants in check—at least as much as for economic reasons.

The League of Nations mandate

The League of Nations awarded a Class C mandate (meaning no real targets for development of the people toward independence were intended) to the crown of Great Britain to be exercised by the Union of South Africa authorities. That "sacred trust" was read as justifying settlement, greater exploitation, and no rights for black (and precious few for Coloured) Africans, plus a creeping annexation into South Africa as a "fifth province." The rail system was extended to Walvis Bay (the one good natural port) and south to the South African border and to Capetown to tie South West Africa's economy to South Africa's on both the import and export sides.

South Africa extended direct rule to the Kunene and Okavango rivers—parallel to a Portuguese push south to the Angola-Namibia border. Resistance there and elsewhere in South West Africa flared into violence repeatedly until the 1930s, while trade union organizing and political as well as economic resistance began in the 1920s. Until 1945 South West Africa was not a productive colony—cattle and karakul were in oversupply, diamond output was held low, and export prices for base metals were not attractive. Governance, security, and settler survival all had to be financed in large part from Pretoria.

The political economy of a colonial boom. From 1945 the economy of South West Africa grew rapidly, reaching a peak of more than $1,000 per capita ($20,000 for Europeans and $150 for black Namibians) in the late 1970s. The pillars were base metal expansion at better prices, sharply increased output and prices for cattle (largely in South Africa), karakul (via South Africa to the European–North American fur market), and diamonds. Fourfold growth in world demand after World War II led to increases in output at De Beer's diamond mines. In addition, the fish catch (largely for fish meal and canned pilchards) exploded to 1,102,000 short tons (1,000,000 metric tons)—a level that laid the groundwork for the present stock depletion and conservation problems.

The economic gap

The European enclave boomed. The situation was quite different for the other 90 percent of the people. Rising population was eroding productive capacity—per capita and absolutely by ecological damage—in African areas. Until the late 1970s, contract labour paid only enough to support a single person at subsistence level. Black nurses, teachers, and secretaries, as well as semiskilled workers, began to be trained and employed on a significant scale only in the mid-1970s. Land reallocations increased contract labour. A body called the Odendaal Commission organized separate development, which led to the creation of "homeland" authorities that benefited a new black elite (as in the 1980s did government wages and salaries for teachers, nurses, and black-area administrators and troops and a wage increase by large employers in mining and finance). A rising proportion of black Namibians—two-thirds by the late 1980s—was left in abject poverty. Further, contract labour eroded the social and civil structures, giving rise to numerous usually very poor female-headed households in the "homelands" and the urban peripheries.

From resistance to liberation struggle. From 1947, Namibians (initially via intermediaries) had begun to petition the United Nations against South African rule. A series of cases before the International Court of Justice (World Court)—the last, in 1971, declaring the mandate forfeiture by the United Nations in 1966 to be valid—led to a de jure UN assumption of sovereignty and de facto support via publicity, negotiation, and training for Namibian liberation.

In South West Africa the churches (numbering at least 80 percent of black Namibians in their membership) took an early lead in petitioning the UN and South Africa and created a climate of black social and civil opinion favourable to the liberation struggle; they were slow, however, to endorse its armed phase. From the 1950s to the 1970s the churches had become increasingly national in staff and outlook, in some cases after severe conflicts with the overseas "parent" bodies and local missionaries.

Black trade union activity (illegal until the mid-1980s) began to revive as well and focused rather more on political than on economic mobilization. The major strike of 1971–72 was against contract labour, the implementation of apartheid, and the 1966 failure of the initial World Court case as much as it was for wage increases per se.

Formation of SWAPO

From 1958 to 1960 the political focus turned from resistance to liberation, and leadership passed from traditional chiefs to party leaders. SWAPO (nominally South West Africa People's Organization, although only the acronym has been used since 1980) was founded as the Ovamboland People's Organization in 1958; it achieved a national following as SWAPO in 1960. In 1959 SWANU (South West Africa National Union) was formed, largely

by Herero intellectuals. Within a decade, SWAPO had become the dominant party and had grown beyond its Ovambo roots. The presence of Ovambo throughout the nation due to contract labour was used to forge a national communication system and mobilizing capacity.

The parties had been formed because petitioning seemed ineffective. The forced removal (with violence and deaths) of black Namibians from the Old Location in Windhoek to the outlying township of Katatura (sometimes translated as "The Place We Do Want to Be") was perhaps the key catalytic event. Until 1966 the parties sought—in the face of increasing repression—to press for redress of grievances from South Africa and via the United Nations. Indeed, until the 1970s the armed struggle, then largely across the border from Zambia, was only a minor nuisance to South Africa.

The 1971–72 strike marked a turning point in terms of national solidarity and nationwide participation in the struggle. It greatly alarmed South Africa; a rising crescendo of trials and summary imprisonment and torture was pursued, though this process had already begun when Herman Toivo ja Toivo and most other SWAPO leaders not already in exile were tried for terrorism and sent to Robben Island in 1968. From 1969 SWAPO had operated along almost all of the northern border—an operation that was easier after Angolan independence in 1975—and in the north-central farming areas around Grootfontein. Although set back by an internal leadership crisis and division among fighting cadres in 1976, the armed struggle had become militarily damaging and economically costly to South Africa by the end of the 1970s.

The road to Namibia. From 1977 through 1988, the economy of Namibia stagnated overall and fell by more than 3 percent per year per capita. Five factors influenced this: six years of drought, decline in fishing yields because of overfishing, serious worsening of import-export price ratios, the slow growth and mismanagement of the South African economy, and the impact of the war on the budget and on both domestic and foreign investor confidence. For white residents, real incomes (except in ranching) stagnated or rose only slowly; for blacks, they rose for perhaps one-sixth of households in wage employment with government or large enterprises and declined rapidly for others, especially for residents of the northern "operational area" (war zone).

For South Africa, Namibia turned from an economic asset to a millstone (with a war bill by the late 1980s on the order of $1 billion a year—comparable to Namibia's gross domestic product). Capital stock was run down, and output of all major products—beef, karakul, fish, base metals, uranium oxide, and diamonds—fell.

On the domestic side a long series of South African attempts to build up pro-South African parties with substantial black support failed even when trade unions were legalized, wages raised, and petty apartheid laws (including abolition of the contract labour and residence restrictions) relaxed. Indeed, after the failure of the Smith-Muzorewa alliance in the Zimbabwe independence elections, South Africa's internal political maneuvers looked increasingly desperate and lacking in conviction.

International response

Internationally and militarily, decline was slower and less apparent. While the UN had passed Security Council resolutions (notably resolution 435) demanding independence, South Africa skillfully protracted negotiations and played on U.S. fears of communism and paranoia about Cuba (whose troops had defeated the 1975 South African invasion of Angola and remained there to augment the defense against South Africa and its Angolan allies or proxies) to cause repeated, time-consuming stalemates.

Through 1986 about 2,500 South African soldiers had died, a figure proportionally higher per capita than the U.S. death toll in Vietnam. However, the South African government skillfully disguised the high casualty rate as well as the fiscal burden of the Namibian occupation and the forward policy in Angola. The war, like the negotiations, appeared stalemated.

The turning point came in 1988. South Africa's invasion of Angola was defeated near Cuito-Cuanavale; air control was lost; and the Western Front defenses were tumbled back to the border (by a force consisting largely of units of SWAPO's People's Liberation Army of Namibia [PLAN] under Angolan command). By June, South Africa had had to negotiate a total withdrawal from Angola in order to avoid a military disaster, and by the end of December it had negotiated a UN-supervised transition to elections, a new constitution, and independence for Namibia.

(R.H.Gr.)

Independence. The United Nations Transition Assistance Group (UNTAG) opened operations in April 1989. After a disastrous start—in which South African forces massacred PLAN forces seeking to report to UNTAG to be confined to designated areas—UNTAG slowly gained control over the registration and electoral process in most areas.

The 1989 election

The election of 1989, held under the auspices of the UN, gave SWAPO 57 percent of the vote and 60 percent of the seats. Sam Nujoma, the longtime leader of SWAPO, became president. With two-thirds majorities needed to draft and adopt a constitution, some measure of reconciliation was necessary to avoid deadlock. In fact, SWAPO and the business community—as well as many settlers—wanted a climate of national reconciliation in order to achieve a relatively peaceful initial independence period.

As a result, a constitution emphasizing human, civil, and property rights was adopted unanimously by the end of 1990, and reconciliation with settlers and (to a degree) with South Africa became the dominant mood. For the new government, the costs of reconciliation included retaining about 15,000 unneeded white civil servants, deferring the landownership and mineral-company terms issues, and offering de facto amnesty for all pre-independence acts of violence (including those of SWAPO against suspected spies and dissidents in Angola in the late 1980s); the benefits were the takeover of a functioning public administration and economy (with growth rising to 3 percent in 1990) and grudging but real South African cooperation on fishing and use of Walvis Bay. Above all, South Africa forebore from mounting destabilization measures or creating proxy armed forces.

On March 21, 1990, the South African flag was lowered and Namibia's raised at the National Stadium; Namibia subsequently joined the Commonwealth, the United Nations, and the Organization of African Unity. Diplomatic relations were established with many countries. The Namibian Defense Force—which included members of PLAN as well as the former South West African Territory Force—was created with the assistance of British military advisers.

South Africa agreed to a transition to Namibian sovereignty over Walvis Bay, which was effected in 1994. It also agreed to a revised boundary along the Orange River, giving Namibia riparian rights; the earlier border had been placed on the north bank and thus left Namibia without water rights. Namibia remained a member of the Southern African Customs Union.

The political climate remained calm. The main opposition party, the Democratic Turnhalle Alliance (heir to South Africa's puppet government efforts and beneficiary of considerable South African funds for campaigning), held almost one-third of the seats in the legislature but was neither particularly constructive nor totally obstructive. In the 1994 national elections, SWAPO consolidated its hold on power, surpassing the two-thirds majority needed to revise the constitution.

SWAPO maintained its hold on power in the country's 1999 elections, despite allegations from the opposition—now headed by a SWAPO splinter party, the Congress of Democrats—that the government was engaging in authoritarian practices. While opponents questioned several of the government's decisions—its support of the Congolese (Kinshasa) government by sending troops into the country in 1998 and its 1999 support of the Angolan government when it allowed government troops to pursue Angolan rebels into Namibian territory—other internal issues continued to be significant, such as Namibia's increasing number of AIDS cases and its continuing debate over the question of land reform. (R.H.Gr./Ed.)

For later developments in the history of Namibia, see the BRITANNICA BOOK OF THE YEAR.

South Africa

The Republic of South Africa, the southernmost state on the African continent, has an area of 470,693 square miles (1,219,090 square kilometres). It measures almost 1,000 miles (1,600 kilometres) from north to south, as well as from east to west. South Africa is bordered by Namibia to the northwest, by Botswana and Zimbabwe to the north, and by Mozambique and Swaziland to the northeast and east. Lesotho, an independent constitutional monarchy, is entirely surrounded by South African territory in the eastern part of the republic. South Africa's coastlines border the Indian Ocean to the southeast and the Atlantic Ocean to the southwest. The capitals are Pretoria (executive), Cape Town (legislative), and Bloemfontein (judicial).

South Africa is relatively isolated, distant even from major African cities such as Nairobi, Kenya (more than 1,500 miles away), and Lagos, Nigeria (more than 2,400 miles away); it is 5,100 miles from South America, 4,700 miles from Australia, and more than 6,000 miles from most of Europe, North America, and eastern Asia, where many of its major economic links lie.

The four original provinces of South Africa—Cape of Good Hope, Orange Free State, Transvaal, and Natal—were reorganized in 1994 into nine new provinces: Western Cape, Northern Cape, Eastern, North-West, Free State, Gauteng, Eastern Transvaal, Northern, and KwaZulu/Natal; Eastern Transvaal subsequently was renamed Mpumalanga. South Africa possesses two small subantarctic islands, Prince Edward and Marion, situated in the Indian Ocean about 1,200 miles southeast of Cape Town. The former South African possession of Walvis Bay, an enclave on the Atlantic coast some 400 miles north of the Orange River, was transferred to Namibia in 1994.

South Africa has long been a focus of world attention. The former South African government, dominated by the minority white population, maintained a policy of apartheid ("apartness") that enforced segregation between government-defined races in housing, education, and many other spheres of life. Apartheid evoked vehement opposition internally and from most countries in the world. In 1990 the South African government began repealing the apartheid laws, initiating the transition to government led by the black majority. This process culminated in the permanent nonracial constitition promulgated in 1997.

(A.Ne./D.F.G./A.S.Ma.)

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* A plateau that covers the largest part of the country dominates the topography. It is separated from surrounding areas of generally lower elevation by the Great Escarpment. The plateau consists almost entirely of very old rock of the Karoo (Karroo) System, which formed from the Late Carboniferous Epoch (320 to 286 million years ago) to the Late Triassic Epoch (230 to 208 million years ago). The plateau is generally highest in the east, dropping from elevations of 8,000 feet (2,440 metres) in the basaltic Lesotho region to 2,000 feet in the sandy Kalahari in the west. The central part of the plateau comprises the Highveld, which is between 4,000 and 6,000 feet in elevation. South of the Orange River lies the Great Karoo region.

The Great Escarpment

The Great Escarpment, known by a variety of local names, forms the longest continuous topographic feature in South Africa and provides scenery of great beauty. It runs southward from the far northeast, where it is generally known as the Transvaal Drakensberg (*berg* and *berge* in Afrikaans means "mountains"). Farther south it forms the boundary first between KwaZulu/Natal and Free State and then between KwaZulu/Natal and Lesotho. There it reaches heights up to nearly 11,000 feet, including some of the country's highest peaks, such as Mont aux Sources (10,823 feet [3,299 metres]); it is known both as Khahlamba (in Zulu) and as the Natal Drakensberg. The mountainous escarpment continues southwestward, dividing Lesotho from the Transkei region of Eastern province. At lesser altitudes of 5,000 to 8,000 feet, it runs westward across Eastern province, where it is known as the Stormberg. Farther to the west, with names such as the Nuweveld Range and the Roggeveld Mountains, it forms the approximate boundary between Northern Cape and Western Cape provinces. At its western extreme, in the vicinity of Mount Bokkeveld and Mount Kamies (5,600 feet), the escarpment is not well defined.

An area of very old folded mountains with altitudes between 3,000 and 7,600 feet lies in the southwest of the country; it includes ranges such as the Tsitsikama, Outeniqua, Great Swart, Lange, Seder, Drakenstein, and Hottentots Holland Mountains, as well as Table Mountain and its associated features at Cape Town.

Both above and below the Great Escarpment, the topography tends to be relatively broken. Open plains are relatively rare, occurring mainly in northwestern Free State and farther to the west and in smaller areas such as the Springbok Flats north of Pretoria. Ridges, mountains, and deeply incised valleys are common, mainly left by the erosion of very old landforms. Between the escarpment and the sea there is little genuine coastal plain, with exceptions in northern KwaZulu/Natal, where it reaches 50 miles in width, and in parts of Western Cape. For most of its 1,836-mile length, the coastline is characterized by fairly steep slopes rising rapidly inland. Most of the coastline has experienced uplift or falling sea levels in the recent geologic past, with the result that few flooded river valleys or natural harbours occur. Exceptions include the Knysna Lagoon in Western Cape and the Buffalo River at East London. Long stretches of beach are common. In KwaZulu/Natal, longshore drift over many centuries has created spits and bluffs from beach sand; in a number of places these features have enclosed bays, which have provided both remarkable sanctuaries for wildlife (as at the St. Lucia estuary) and, when mouths are dredged, good harbours (as at Durban and Richard's Bay).

Drainage. The greater part of the country (about 329,000 square miles) is drained to the Atlantic Ocean by the Orange River, which rises in the Lesotho Highlands, and its tributaries. Chief among the latter are the Caledon and the Vaal. North of the Witwatersrand ridge, the plateau is drained to the Indian Ocean by the Limpopo system, with major tributaries in the Krokodil, Mogalakwena, Luvuvhu, and Olifants rivers. South of the Olifants River, the area between the escarpment and the sea is drained by a large number of other river systems, such as the Komati, Pongola, Mfolozi, and Mgeni, the largest of which is the Tugela, draining much of KwaZulu/Natal and ranking as the largest river by volume in the country. To the south the Mkhomazi, Mzimvubu, Great Kei, Great Fish, Sundays, and Gourits rivers drain significant areas; the Western Cape fold mountain region is drained mainly by the Breë, Berg, and Olifants rivers. All South African rivers are highly seasonal in flow, and few offer enough gradient and volume to allow navigation by even small craft for more than a few miles from the river mouths.

Orange River drainage pattern

Soils. Three major soil regions may be distinguished. East of approximately longitude 25° E, soils have formed under wet summer and dry winter conditions, the more important types being laterite (red, leached, iron-bearing soil) and other lateritic soils, unleached subtropical soils, and gleylike (*i.e.,* bluish gray, sticky, and compact) podzolic soils (highly leached soils that are low in iron and lime). A second major region lies within winter and all-season rainfall areas in Eastern and Western Cape; its soils are generally gray sandy and sandy loam in type. Over most of the rest of the country, which is generally dry, soils are characterized by a sandy top layer, often a sandy loam, underlain by a layer of lime or an accretion of silica. With some exceptions, South Africa's soils are not characterized by high fertility, and those that are—for example, in coastal KwaZulu/Natal—tend to be easily degraded.

Climate. The climate is varied but affected by three main factors. First, South Africa's location between latitudes 22° and 35° S places almost the entire country within the temperate zone; extremes of heat and cold are rare. However, its location next to a subtropical high-pressure belt of descending air produces stable conditions unconducive to rainfall over most of its surface. The result is a generally dry climate.

Second, most of the country lies at fairly high elevations,

which tempers the influence of latitude and makes even the tropical and near-tropical inland areas much cooler than would otherwise be the case. High elevation and lack of the moderating influence of the sea give most inland areas large daily variations in temperature.

Third, the climate is greatly influenced by the ocean that surrounds the country to the east, south, and west. The temperate cyclones of the southern ocean exercise considerable influence on weather patterns, especially in winter, when their circulation moves northward. The cold, northward-flowing Benguela Current not only cools the west coast considerably but also contributes to the dryness and stability of the atmosphere over the western parts of the country, while the warm, southward-flowing Mozambique and Agulhas currents keep temperatures higher on the east and southeast coast. The resultant warmer and less dense air rises more readily, facilitating the entry of rain-bearing clouds from the east.

Influences of the subtropical high-pressure belt

South Africa, as well as the adjoining ocean areas, is influenced throughout the year by descending, divergent upper air masses that circulate primarily eastward, generally causing fine weather and low annual precipitation, especially to the west. During winter (June to August), cold polar air moves over the southwestern, southern, and southeastern coastal areas, sometimes reaching the southern interior of the country from the southwest. These polar masses are accompanied by cold fronts as well as by rain and snow. In summer (December to February), the Atlantic high-pressure system settles semipermanently over the southern and western parts of the country. Local heating of the landmass sometimes causes low-pressure conditions to develop, resulting in tropical air masses being drawn in from the Indian Ocean over the northeastern region and bringing rain.

South Africa is a generally semiarid country where farmers often face water shortages. More than one-fifth of the country is arid, receiving less than 8 inches (200 millimetres) of rainfall annually; almost half is semiarid, receiving 8 to 24 inches. In most of these areas, rainfall is highly variable and generally unreliable. Only 6 percent of the country receives more than 40 inches of rain annually. From east to west there is a gradual decline in the rainfall. The KwaZulu/Natal coast receives more than 40 inches annually, Kimberley approximately 16 inches, and Alexander Bay on the west coast less than 2 inches.

Summers are warm to hot, with daytime temperatures generally from 70° to 90° F (21° to 32° C). Higher elevations have the lower temperatures, while the far northern and northeastern regions and the western plateau and river valleys in the central and southern regions have the higher temperatures. At night, temperatures fall substantially in the interior—in some places by as much as 30° F (17° C)—while on the coast the daily range is much smaller. Winters are mostly cool to cold, with many high-elevation areas often having temperatures below freezing at night but readings of 50° to 70° F (10° to 21° C) in the daytime. On the eastern and southeastern coast, winters are warm. Temperatures generally decline from east to west: Durban has an annual average temperature of 69° F (21° C), while Port Nolloth—at a similar latitude but on the west coast—registers 57° F (14° C).

Plant and animal life. Natural vegetation varies from savanna (parklike grassland) in the Bushveld and Lowveld of Mpumalanga and Northern provinces through grassland with fewer trees in the Highveld to scrub and scattered bush in the Karoo and drier western areas. Western Cape has a distinct vegetation of grasses, shrubs, and trees able to withstand the long, dry summers and is the home of many of South Africa's 20,000 species of flowering plants. The eastern coast has a more tropical plant life. Natural forest is limited to mountainous valleys along the Great Escarpment and a few other favoured localities.

South Africa has a rich and varied mammal life, with more than 200 species, including large animals such as lions, leopards, elephants, rhinoceroses, hippopotamuses, baboons, and many kinds of antelope. Smaller creatures include mongooses, jackals, and various cats. However, the numbers of animals declined greatly during the expansion of white settlement in the 18th and 19th centuries, and today large indigenous mammals occur mainly in wildlife reserves, of which Kruger National Park is the best known. There are more than 800 species of birds, many species of reptiles, including more than 100 snake species, and an extraordinary diversity of insects. (A.Ne./A.S.Ma.)

South Africa contains more than a dozen national parks, including Kgalagadi Transfrontier Park, formerly Kalahari Gemsbok Park, which was joined with Botswana's Gemsbok National Park to create a transnational park. The largest, Kruger National Park in Northern and Mpumalanga provinces, is noted for its populations of rhinoceroses, elephants, and buffalo. Mountain Zebra National Park in Eastern province shelters the endangered mountain zebra; Addo Elephant National Park, also in Eastern, protects more of the elephant population; and Bontebok National Park in Western Cape contains the endangered bontebok (a type of antelope). Regulated big-game hunting is allowed in the country from May until July for elephants, white rhinoceroses, lions, leopards, buffalo, and many types of antelope. Grysbok, klipspringers, and red hartebeests (all of them varieties of antelope), giraffes, black rhinoceroses, pangolins (anteaters), and antbears are specially protected animals that cannot be hunted. (Ed.)

Settlement patterns. More than nine-tenths of the people live in the eastern half of the country and in the southern coastal regions. The western parts, except for the area around Cape Town in the extreme southwest, are very sparsely populated. More than half the population lives in urban areas, and many of those live in or around the major cities. About half of the black population is concentrated in the former "homeland" areas, scattered territories in the northern and eastern parts of the country that were established during the apartheid period as segregated living areas for Africans. Far from urban facilities, some of these areas exhibit urban, rather than rural, population densities.

White conquest and changing land tenure

Rural land is predominantly owned today by whites but was historically settled by blacks in much of the country. Traditional African settlements consisted of commonly owned farming homesteads or villages (kraals). Conquest and the establishment of white authority and private ownership of land made these settlement patterns subordinate to others, and in the modern period about one-sixth of the African population came to live on farmland owned by whites.

Rural patterns created by white settlement from the late 17th century differ in being centred around single homesteads, usually considerable distances apart, each having its associated cluster of sharecroppers', tenants', or employees' houses. As the frontier of white settlement expanded in the 18th and 19th centuries, the establishment of farms often involved each farmer claiming several thousand acres, resulting in a dispersed settlement pattern with homesteads sometimes miles apart. Smaller farms and more intensive cultivation, however, always existed in some areas, such as the grape-growing areas of the southwest. From the late 19th century, as the urban demand for food and other agricultural produce grew rapidly, many farms closer to towns or in more favourable ecological zones were subdivided, and a denser pattern emerged. More recently there has been a general tendency for farm sizes to increase as the number of landowners has declined.

Urban settlement in South Africa originated both with the concentration of population around the political centres of African chiefdoms and kingdoms and with the towns established by European colonizers. The general defeat of African polities by whites and their allies, particularly during the 19th century, led to the abandonment or destruction of capitals such as Dithakong, a Tswana stronghold in what is now Northern Cape, and Ulundi, the Zulu citadel in northern Natal (now KwaZulu/Natal).

European colonization of South Africa began with towns, Cape Town being the first, in 1652. Under Dutch rule, a few Dutch colonial towns were established in the south and southwest, including Stellenbosch, Tulbagh, Graaff-Reinet, and Swellendam. Once British rule began at the start of the 19th century, new towns were created more rapidly, such as Port Elizabeth, Grahamstown, Beaufort West, and Durban. The Great Trek of dissatisfied farmers

and townspeople away from British rule, which commenced during the 1830s, led to a range of new, mainly small urban centres in the interior focused on church and government: such places included Winburg, Pietermaritzburg, Potchefstroom, Bloemfontein, and Pretoria.

Cape Town

Until the 1860s all South African towns were small; the largest, Cape Town, had a population of less than 40,000 in 1865. Urbanization accelerated rapidly from the 1870s as railway building, mining, and economic expansion proceeded. The population of the Cape Town metropolitan area reached 130,000 by the beginning of the 20th century, by which time Johannesburg (established in 1886) had surpassed its size. Continued rapid growth since then has created four major urban concentrations. Of these by far the largest is the Pretoria-Witwatersrand-Vereeniging complex, centred on Johannesburg, radiating about 45 miles in each direction, and now mostly falling into Gauteng province. The Durban area forms the second largest urban region, followed by that centred on Cape Town and the Port Elizabeth–Uitenhage area in Eastern province, the smallest of the major industrial centres. The major centres in these metropolitan areas offer the full range of services that cities of their size offer in other countries, but all show great disparities of income and access to urban services between the wealthiest, predominantly white, areas and the poorest, exclusively black, districts.

Outside these major metropolitan areas, most South African towns are small and serve either mining communities or surrounding rural areas. Between these extremes are several cities with rapidly growing populations numbering in the hundreds of thousands: the port of East London, the KwaZulu/Natal capital of Ulundi, the Free State capital of Bloemfontein, newer industrial centres such as Witbank in Mpumalanga, and a few rural service centres that have recently become regional administrative and educational centres and grown rapidly, such as Mafikeng, Nelspruit, and Pietersburg.

South African cities have shown a measure of racial segregation in residence since their colonial foundation. Segregated public housing areas were created in the early 20th century. Various measures introduced from the 1920s on gave authorities powers to segregate Africans and others; during the 1930s and '40s, attempts were made to extend such provisions to segregate Coloureds (persons classified as being of mixed race) and Indians from whites as well, culminating in the Group Areas Act of 1950. Under its provisions, South African cities acquired their characteristic form, with white residential areas, generally situated in more favourable localities, occupying most of the urban space. Other sectors and peripheral localities were set aside for those classified as nonwhites; many of those areas were initially devoted to segregated public housing estates called "townships." More recently a degree of racial integration of residence has occurred in some cities; for example, high-density residential areas such as Hillbrow in Johannesburg became effectively integrated despite the Group Areas Act. The act was repealed in 1991 but the racially defined settlement patterns in the towns and townships have persisted.

The people. *Ethnic distribution.* Government-determined racial and ethnic classification, embodied in the Population Registration Act in effect from 1950 to 1991, was crucial in determining the status of South Africans under apartheid. The act divided South Africans at birth into four racial categories—black, white, Coloured, and Asian—though these classifications were largely arbitrary, based on considerations such as family background and cultural acceptance as well as on appearance.

The original Khoikhoi and San peoples of South Africa scarcely exist as distinct groups inside the country today. Many intermarried with other African peoples who arrived before white conquest, and others intermarried with Malagasy and Southeast Asian slaves under white rule to form the backbone of the Coloured population. Bantu-speaking Africans entered the area roughly 1,800 years ago; their descendants today constitute about three-fourths of South Africa's population. The African population is heterogeneous, falling mainly into four linguistic categories. The largest is the Nguni, including various Ndebele-, Swati-, Xhosa-, and Zulu-speaking peoples, who constitute more than half the African population of the country and form the majority in many eastern and coastal regions as well as in the industrial Gauteng province. The second largest is Sotho-Tswana, which includes numerous Sotho-, Pedi-, and Tswana-speaking peoples and constitutes a majority in many Highveld areas. The other two are the Tsonga (or Shangaan) speakers, concentrated in Northern and Mpumalanga provinces, and the Venda speakers, located largely in Northern province.

Major African groups

White South Africans consist of two main language groups. More than half of them are Afrikaans speakers, the descendants of mostly Dutch, French, and German settlers. Most of the remainder are English-speaking, mainly the descendants of British colonists, with a sizable minority of Portuguese origin and smaller groups of Italians and others. Immigration from Europe exceeded 20,000 people per year during the late 1960s and early '70s, but in the late '70s and '80s the number of whites leaving South Africa tended to exceed the new arrivals. The white population is highly urbanized.

The population formerly classified as Coloured descended from Khoisan (Khoikhoi and San) peoples, slaves imported by the Dutch from Madagascar and what is now Malaysia and Indonesia, Europeans, and Bantu-speaking Africans. Several distinct subethnic groups can still be

Rural African settlement in the Nelspruit district of Mpumalanga province, northeastern South Africa, southwest of Kruger National Park.

Mark Segal/Tony Stone Worldwide

identified, such as the Malays, who largely originated from Indonesian Muslim slaves, and the Griquas, who trace their origins to a specific historical Khoikhoi community. While some Malays and Griquas have continued to identify themselves as Coloured, others who were so classified by the apartheid government have tried to reject the label entirely. Most of this population speaks Afrikaans or, to a lesser extent, English, and in many respects they cannot be distinguished culturally or physically from the white population. Those formerly classified as Coloured are concentrated in the western half of the country, particularly in Western Cape, Northern Cape, and the westernmost parts of Eastern province, where they form a majority in most districts.

South Africans of Indian descent, classified under apartheid as Asian, form a large minority, which originated with immigrant traders and indentured workers in the second half of the 19th century. Today the large majority live in KwaZulu/Natal and to a lesser extent in Gauteng, Northern, and Mpumalanga. Almost all Indian South Africans live in urban areas. Small communities of other ethnic Asians, including Chinese, live in some of the cities.

Religion. The largest category of religious affiliation is to independent African Christian churches. The largest established Christian denominations, drawing members from all ethnic groups, are the Methodist, Roman Catholic, Anglican, and Dutch Reformed churches. The other major religions are Hinduism, among the majority of Indians; Islām, among many Indians and Malays; and Judaism, among a significant minority of the white population.

(A.Ne./D.F.G./A.S.Ma./Ed.)

The economy. South Africa's economy was revolutionized in the late 19th century when diamonds and gold were discovered there. Extensive investment from foreign capital followed. In the years since World War II the country established a well-developed manufacturing base, and it experienced highly variable growth rates, including some years when its growth rate was among the highest in the world. From the late 1970s South Africa has had continuing economic problems, initially because its apartheid policies led many nations to withhold foreign investment and to impose increasingly severe sanctions against it. South Africa's economy did not immediately rebound in the early 1990s while apartheid was being dismantled, as investors waited to see what would happen. Only since democratic elections in 1994 has significant investment returned. Postapartheid South Africa now faces the problem of integrating the previously disenfranchised and oppressed majority into the economy. The government created a new plan in 1996 called Growth, Employment, and Redistribution (GEAR), focusing on privatization and removing exchange controls.

State participation

The economy is essentially based on private enterprise, but the state participates in many ways. Through the Industrial Development Corporation, the former government set up and controlled a wide array of public corporations, many relating to industrial infrastructure. The Iron and Steel Corporation (ISCOR), the major iron and steel producer, and the South African Coal, Oil, and Gas Corporation (SASOL), which produces oil from coal, were both privatized in the 1980s. The Electrical Supply Commission (ESKOM) remains government-controlled, but in the 1990s the government partially privatized airlines and telecommunications, and despite fierce opposition from trade unions, official economic policy has been to continue partially or completely privatizing many public enterprises.

By means of a range of official and quasi-governmental bodies, the government encourages industrial development, consultations on tariff protection, and export promotion and research; it also maintains a bureau of standards. The Development Bank of Southern Africa is a quasi-governmental company created to promote development projects, while the South African Housing Trust is a joint venture between the government and the private sector intended to provide low-cost housing.

Central government taxation consists primarily of income taxes on individuals and businesses and a value-added tax on transactions. Provincial governments depend mainly on transfer payments from the central government, while property taxes and levies on businesses provide the main support for local governments.

Historically, the stated policy of the African National Congress (ANC) was that it would seek a state-led mixed economy based on nationalized mining and financial enterprises, yet since taking leadership of the government in 1994 it has in fact pursued privatizing a substantial number of formerly state-owned enterprises. The government faces competing demands—to improve the living conditions of the impoverished black population while also addressing the demands for economic liberalization from business interests and Western governments. It has chosen to make maintaining business confidence and boosting investment the core element of its economic policy. (Ed.)

Resources. South Africa is rich in a variety of minerals. In addition to diamonds and gold, the country also contains reserves of iron ore, platinum, manganese, chromium, copper, uranium, silver, beryllium, and titanium. No commercially exploitable deposits of petroleum have been found, but natural gas in moderate quantities exists off the southern coast, and oil is made from coal at two large plants in the provinces of Free State and Mpumalanga.

Diamonds and gold

Large coal deposits exist, mostly at easily mined depths beneath the Mpumalanga and northern Free State Highveld. Vast deposits of platinum-group and chromium minerals are located mainly to the north of Pretoria. There are major deposits of iron ore and manganese, particularly in Northern Cape; titanium-bearing sands are common on the eastern seaboard.

Agriculture is of major importance to South Africa, especially as an employer, but land and water resources are generally poor. There are areas of exception, however, such as the well-watered, fertile soils of the Western Cape river valleys, the KwaZulu/Natal coast, and the Highveld of Mpumalanga and Free State, which offer good conditions for extensive cereal cultivation. There are also some dry areas in which irrigation allows the soils to become productive, such as in the Fish River valley of Eastern province. The Orange River Project is designed to add another three-tenths to the total land in production.

Timber resources are minimal, but the small percentage of land under indigenous forest has been supplemented by substantial areas under plantation in the wetter parts of the east and southeast. Hydroelectric potential is limited, though the government has developed projects on a number of rivers; more significant are the projects to import electricity from stations on the Zambezi at Cahora Bassa, Mozambique, and on rivers in the Lesotho Highlands.

Agriculture, forestry, and fishing. Arable land constitutes slightly more than one-tenth of the country's surface area, but in general agriculture contributes a substantial amount to both the domestic economy and exports. Among the major crops are corn (maize), wheat, sugarcane, sorghum, peanuts (groundnuts), citrus and other fruits, and tobacco. Sheep, goats, cattle, and pigs are raised for food and other products; wool and meat (beef, sheep, and goat) are important. Dairy (including butter and cheese) and egg production are also significant, particularly around the major urban centres.

Major agricultural products

The forest industry supplies mining timber, pulpwood for paper and board mills, and building timbers mostly sufficient for a construction industry that primarily uses brick, concrete, and steel. Fishing areas lie mainly off the western and southern coasts. The principal shoal-fishing products are pilchard and maasbanker, while offshore trawling produces kingklip, Agulha sole, Cape hake, and kabeljou, among others.

Industry. Although for decades manufacturing has employed more people and produced more output in terms of percentage of the GDP than has mining, the mining sector continues to form the core of the South African economy. Gold remains the most important mineral—South Africa is the largest producer in the world—and reserves are large; however, production is slowly declining, and prices have not ever equaled their spectacular highs of the early 1970s, rendering a number of older mines marginal or unprofitable. Several gold mines closed in the 1990s, and thousands of mine workers lost their jobs. The main

goldfields centred historically on Johannesburg; the major areas of production now lie some distance east, west (Far West Rand), and south (northern Free State) of Johannesburg, centred on towns such as Evander, Carletonville, and Welkom, respectively.

Coal is South Africa's second most valuable mineral product. The largest sales of coal are for export (to East Asia and Europe) and for the generation of electricity. Platinum and chromium, of which South Africa is the world's largest producer, are produced at centres such as Rustenburg and Steelpoort in the northeast and are becoming increasingly significant. Gem diamond production historically was concentrated around Kimberley but now occurs in a variety of localities. South Africa is among the world's largest diamond producers. Most South African diamond production is controlled by De Beers Consolidated Mines, Ltd.

Electricity is produced thermally almost entirely from coal. Most electric power is generated by ESKOM at huge stations in the Mpumalanga Highveld. Oil derived from coal supplies a percentage of the country's energy needs, the remainder coming from imports refined at the ports or piped to a major inland refinery at Sasolburg.

The major manufacturing sectors are in food processing, textiles, metals, and chemicals. Agriculture and fisheries provide the basis for substantial activity in meat, fish, and fruit canning, the sugar industry, and other processing; more than half the products are exported. A large and complex chemical industry developed from early beginnings in the manufacture of explosives for use in mining. A coal-based petrochemical industry produces a wide range of plastics, resins, and industrial chemicals. The metal industry, centred in Gauteng, draws much of its raw material from ISCOR, the iron and steel producer. Other firms also supply the steel, heavy engineering, and machinery industries. Imported materials supply aluminum manufacturers located mainly in KwaZulu/Natal. Manufacturing extends to automobiles, shipbuilding, building materials, electronics, and many other sectors, notably armaments, though the weapons industry has begun to diversify into nonmilitary production. Manufacturing has been highly dependent on foreign capital; it expanded rapidly in the 1960s and early '70s but grew relatively slowly or even contracted during the '80s. About one-fourth of manufacturing output is exported.

Finance. South Africa has a well-developed financial system. The South African Reserve Bank is the sole note-issuing authority. It formulates and implements monetary policy and manages foreign-exchange transactions. There are more than 50 registered banking institutions, about 15 of which concentrate on commercial banking. There are more than a dozen merchant and discount banks and a number of permanent building societies. The Land Bank is an important specialist state institution, lending to agriculture but historically only to white farmers.

Private pension and provident funds and more than 90 insurance companies play significant roles in the financial sector. There is an active capital market organized around the Johannesburg Stock Exchange.

Trade. Dependence on foreign trade is relatively high, and the South African economy is thus sensitive to global economic conditions. Precious metals and base metals have been leading exports; agricultural goods also play an important role. South Africa exports military equipment worldwide, particularly to the Middle East. The country's major imports are oil, machinery, electrical equipment, and transportation equipment. A number of imports, most importantly oil, have been classified by the government as "strategic materials." Data for these are not available, making it difficult to assess the exact percentages of various import categories. South Africa's main trading partners are the United States, the United Kingdom, Germany, Italy, and Japan. These five countries account for the bulk of the country's trade. However, South Africa's trading partners have diversified; Taiwan, several Latin-American countries, other African countries, and a number of Asian economies such as Singapore, Malaysia, and India account for an increasing proportion of South Africa's international trade. (E.I.U./D.F.G./A.S.Ma.)

Trading partners

Transportation. Because South African rivers are generally unsuitable for navigation, coastal shipping provides the only water transport. Africa's most intense network of roads and railways supplies most of the transportation demand.

The railway system is almost entirely owned and operated by the state; it serves all the major cities, most smaller towns, and many rural areas. A narrow gauge of 3 feet 6 inches (1,067 millimetres) was adopted in the 1870s to lower the cost of construction in mountainous terrain. More than three-fourths of the network of 14,000 miles of track is electrified, and over four-fifths of all traffic is hauled by electric locomotives. More than one-fourth of goods transported consists of coal, other minerals account for one-third, and agricultural products for one-eighth. Long-distance passenger services have declined, but many commuters use train services in all the major urban centres. The luxurious Blue Train—which runs the 1,000 miles between Pretoria, Johannesburg, and Cape Town—and the surviving steam-operated services provide popular tourist attractions.

The road network comprises some 120,000 miles of roads, ranging from rural unpaved stretches to multilane freeways. About 35,000 miles of road are paved. Most towns are connected by high-standard two-lane highways; freeway systems extend around the four major urban areas, but over long distances only Johannesburg and Durban are connected by four-lane highway. Almost all roads are provided and regulated by the different levels of government, but some long-distance roads have been transferred to the private sector, and tolls are charged on them.

Inland air transport services, both passenger and freight, are operated by state-owned South African Airways and by an increasing number of private competitors. Air services connect all major cities. South African Airways and many foreign carriers fly between South Africa and all neighbouring countries; international service extends to six continents. The international airport near Johannesburg is the main hub of South African air transport both domestically and internationally, while the airports at Cape Town and Durban play increasingly important roles as international destinations.

Air services

All South African ports are owned and operated by the government. Durban, which serves most of KwaZulu/Natal, Mpumalanga, and northern Free State, is the major port. Port Elizabeth, Cape Town, and East London handle mixed traffic for their immediate hinterlands and more distant locations. All these ports handle goods traveling to and from other African countries, including Zimbabwe, Zambia, and Zaire. Maputo, in Mozambique, which is the port closest to Johannesburg, serves many areas of the northern provinces. In addition, newer ports have been developed at Richard's Bay on the north coast of KwaZulu/Natal and in the excellent natural harbour at Saldanha Bay north of Cape Town, mainly for exports of coal and iron ore, respectively. (A.Ne./D.F.G./A.S.Ma.)

Administration and social conditions. The republic's original constitution, the South Africa Act of 1909, formed a parliamentary system with the British monarch as head of state. The constitution was revised by the Republic of South Africa Constitution Act of 1961, which transformed South Africa from a dominion within the Commonwealth to an independent republic. South Africa's political development has been shaped by the implementation of apartheid policies by the white minority, the ensuing widespread protest and social unrest, and the adoption in 1993 of a nonracial interim constitution that took effect in 1994. A new, permanent constitution, mandated by the interim document and drafted by Parliament, entered force in 1997.

Government under apartheid. Until 1994 the three officially designated nonwhite groups—Africans, Coloureds (those of mixed race), and Asians (primarily Indians)—were systematically deprived of political participation in the conduct of national and provincial affairs. With few exceptions, the nonwhite population was prohibited from voting. In 1959 African representation in Parliament, which had been provided by three elected whites, was abolished. The 1984 constitution extended the franchise

to Coloureds and Asians and established separate legislative chambers for whites, Coloureds, and Asians. Africans continued to be excluded from the national government.

The constitution. The constitution points to the injustices of South Africa's past in its preamble and defines the republic as a sovereign democratic state founded on the principles of human dignity, nonracialism and nonsexism, and the achievement of equality and advancement of human rights and freedoms. Another of its guiding principles, that of "cooperative government," emphasizes the distinctiveness, interdependence, and interrelationship of the national, provincial, and local spheres of government.

The constitution calls for a bicameral national Parliament. The lower house, or National Assembly, comprises 350 to 400 members who are directly elected to a five-year term through proportional representation. The National Council of Provinces, which replaced the Senate as the upper house, is made up of 10-member delegations (each with six permanent and four special members, including the provincial premier) chosen by each of the provincial assemblies. For most decisions each delegation casts a single vote. The president, elected from among the members of the National Assembly by that body, is the head of state and, as the head of the national executive, presides over a cabinet that includes a deputy president and a member whom the president designates as the "leader of government business" in the assembly. All citizens aged 18 years or older have the right to vote.

Provincial, regional, and local government. The South Africa Act of 1909 created a unitary state in which the four former colonies became provinces governed by a white-elected council. These provincial councils were abolished in 1986, and the executive committees, appointed by the president, became the executive arms of the state in each province. By the late 1980s a small number of Africans, Coloureds, and Indians had been appointed to them.

The constitution of 1996, which abolished the provinces in this form and replaced them with nine new provinces, provides for the election of provincial legislatures comprising 30 to 80 members, elected to five-year terms through proportional representation. Each legislature elects a premier, who then appoints a provincial executive council of up to 10 members. The provincial legislatures have the authority to legislate in a range of matters specified in the constitution, including education, environment, health, housing, police, and transport, although complex provisions give the central government a degree of concurrent power. South Africa thus has a weak federal system.

Urban municipal government has existed in South Africa since the early 19th century but has been unevenly developed. Local councils were elected by white voters only, with rare exceptions; Coloureds and Asians were allowed only advisory powers. In 1990 local government entered a phase of transition, and new arrangements came under negotiation. Under the 1996 constitution local government is predicated on the division of the entire country into municipalities. Executive and legislative authority is vested in municipal councils, some of which share authority with other municipalities. Chiefs (Zulu: amaKhosi; Sotho: Makgosi) were once the hereditary leaders of the various clans or tribes of African people in what is now South Africa. Under colonial rule, attempts were made to rule indirectly through chiefs, and under apartheid the chiefs were incorporated into the government, their appointments being dependent on the governing authorities. Chiefs generally work with traditional councils (known under apartheid as "tribal authorities").

Political parties. The all-white National Party (NP) was the dominant parliamentary party from 1948 to 1994. Before 1990 its programs emphasized white South African nationalism, anticommunism, and the implementation of apartheid. The largest black political organization has long been the African National Congress (ANC), formed in 1912 but banned 1960–90. The Inkatha Freedom Party (IFP, or Inkatha), founded in the mid-1970s, represented many Zulu people.

The ANC and the NP dominated the constitutional negotiations that began in 1990 and included about two dozen other groups. In the 1994 elections, the ANC took more than 60 percent of the vote, the NP about 20 percent, and Inkatha about 10 percent. The ANC won majorities in seven of the nine provinces; the NP in one, Western Cape; and Inkatha won a majority in KwaZulu/Natal. Racially defined voting patterns began to dissolve. The other parties receiving significant support were the Freedom Front (a right-wing white party), the Democratic Party (the heir to a long liberal tradition in white politics), and the Pan-Africanist Congress, a small group that broke away from the ANC in 1959. The South African Communist Party, the South African National Civic Organization, and the trade union federation entered candidates on the ANC's lists.

Provinces of South Africa.

Justice. The common law of the republic is based on Roman-Dutch law, the uncodified law of the Netherlands retained after the Cape's cession to the United Kingdom in 1814. The judiciary comprises the Constitutional Court (with powers to decide on the constitutionality of legislative and administrative actions, particularly with respect to the bill of rights), the Supreme Court of Appeal (the highest court of appeal except in constitutional matters), the High Courts, and Magistrate's Courts. The Supreme Court is headed by a chief justice, who is appointed by the state president, as are the deputy chief justice and the chief justice and deputy chief justice of the Constitutional Court. Other judges are appointed by the president with the advisement of the Judicial Service Commission. Traditional authorities exercise some powers in relation to customary law, which derives from indigenous African practice codified in some areas (such as KwaZulu/Natal) by colonial rulers. Customary law continues to be recognized in various ways. Most civil and criminal litigation is a matter for the Magistrate's Courts.

Armed forces and security. In 1994 the armed forces entered a period of transition. South Africa's military traditionally had been white, with a small standing force and a large reserve component. However, from the 1970s an increasing number of black troops were recruited. Compulsory military service, formerly for white males only, ended in 1994. Integration of forces established as part of the antiapartheid struggle brought change to the military as the guerrillas of the ANC's military wing, Umkhonto we Sizwe ("Spear of the Nation"), were incorporated into the South African army.

Integration of the military

The navy has a small fleet consisting of frigates, submarines, minesweepers, small strike craft, and auxiliary vessels. The air force's craft include fighter bombers, interceptor fighters, helicopters, and reconnaissance, transport, and training aircraft. (M.J.L./D.F.G./A.S.Ma./Ed.)

During the apartheid period the South African government, through a network of private and government-

controlled corporations led by the state-owned Armaments Corporation of South Africa (Armscor), developed a variety of new weapons systems, mostly in order to overcome the effects of the international arms embargo imposed by the United Nations Security Council in 1977. Nuclear weapons were developed in great secrecy—six atomic bombs were built during the 1970s and '80s—but the nuclear weapons program was terminated in 1989, and the bombs were dismantled the following year.

The regular police are organized nationally and comprise regulars as well as reservists. There have been about equal numbers of whites and nonwhites. The police bear the responsibility of maintaining internal security; this brought them into sharp conflict with antiapartheid demonstrators during the 1970s and '80s. Freed of the burden of enforcing apartheid, the police face the challenge of forging better relationships with communities in the fight against rising crime levels.

In the late 1970s the daily average prison population was almost 100,000, one of the highest rates in the world. Of these the majority were imprisonments for statutory offenses against the so-called pass laws, repealed in 1986, which restricted the right of Africans to live and work in white areas and which did not apply to other racial groups. Since then the proportion of the population in prison has declined, as the ending of the state of emergency in 1990 and the process of negotiating a new constitution led to the release of many political prisoners. An amnesty policy was instituted, covering politically inspired offenses committed on both sides during the closing years of the struggle against apartheid.

Education. Under apartheid, different departments of education—segregated by racial category—ran school systems that varied greatly in standards and facilities. Schools run by white education departments had the best resources in the public school system. From 1990 some of these schools began to admit black pupils according to a variety of limited models of change. Since the 1993 constitution outlaws discrimination on grounds of race, schools are now compelled to admit all applicants, but capacity limitations and fees generally have kept Africans out of historically white public schools. Private schools, many of which offer superior educational programs, remain largely inaccessible to Africans because of the high cost.

Educational reform

The great challenge facing the provincial departments of education is to repair the decimated system of African education inherited from the apartheid years. The majority of pupils are in schools formerly run by the old "homeland" governments; these schools were characterized by poorly trained teachers, a corrupt and mismanaged bureaucracy, and a chaotic learning environment. Some schools became centres of protest and rebellion, and students were often absent for long periods because of boycotts. Many African schools were severely overcrowded and lacked basic structural necessities such as indoor plumbing, heat, and electricity; educational materials including textbooks, paper, and desks were in short supply. In 1993 a National Education and Training Forum, comprising representatives of the government and private sector, educators, parents, and students, was established to address the education crisis and develop a comprehensive reform program. School education is compulsory for all children between 7 and 16 years of age.

Higher education

The oldest and largest of the universities is the University of South Africa, which began in Cape Town but is now based in Pretoria and which offers correspondence courses in both English and Afrikaans. The oldest of the residential universities are those of Cape Town, Fort Hare, Stellenbosch, and Witwatersrand (Johannesburg); of these, only Stellenbosch is an Afrikaans-language institution, while Fort Hare was originally established to serve Africans only. Newer Afrikaans institutions are the Universities of Pretoria and Potchefstroom and Rand Afrikaans University (Johannesburg), while the University of Port Elizabeth is bilingual. The English-language institutions, including the University of Natal (Pietermaritzburg and Durban) and Rhodes University, to some extent admitted black students prior to 1959, when their ability to do so was undermined by apartheid legislation that they fiercely opposed. The government then established several new institutions (the Universities of the North, Zululand, Western Cape, Durban-Westville, and Vista and the Medical University) for various black groups. The former homelands of Bophuthatswana, Transkei, and Venda also established their own universities. In 1983 official university apartheid ended, but the various institutions remain influenced by their historically dominant ethnic character. Professional and postgraduate courses are concentrated at the formerly white universities. Technically oriented education is offered by a range of *technikons* and technical colleges.

Health and welfare. While racial bias was not explicitly written into health legislation during the apartheid period, health care for South Africans invariably reflected the economic and political inequalities of the society. Hospital segregation has ended, but access to health services remains greatly inferior in historically black areas. The health status of Africans is generally low; malnutrition is perhaps the most important example, especially among rural children. There is an enormous discrepancy in infant mortality rates, which are lowest for whites and highest among rural Africans. The number of South Africans infected with HIV, the virus that causes AIDS, increased sharply during the 1990s. Since 1994 both the Department of National Health and the administrations of the new provinces have emphasized primary health care delivery, building in some instances on programs that farsighted medical workers instituted during the apartheid period.

In the cities and large towns a highly sophisticated public health system exists. Some of the largest public hospitals are linked to the university medical schools, but those located in the formerly segregated African areas tend to be overcrowded. Many of the more expensive private hospitals are accessible only to those of higher incomes. Most regularly employed persons enjoy a degree of private medical insurance; for many of the more affluent, private general practitioners and specialists supply most needs. National policy development in the postapartheid period may include some form of national health insurance, at least for poorer people.

Government provides a number of welfare measures, among them small pensions for all citizens beyond retirement age whose incomes are below a minimal level. Large numbers of elderly Africans, and often their dependents, gain a minimal livelihood from this system. In the past, welfare systems were administered separately for different defined racial groups; the value of pensions was greatest for whites, less for Indians and Coloureds, and lowest for Africans. During the late 1980s the differentials began to be reduced, and the 1993 constitution prohibits discrimination on grounds of race.

Local authorities have been responsible for public housing since the 1920s; it has been segregated by race, and control over African housing reverted to the central government in 1971. Much of the housing built for Coloureds and Indians was used to rehouse communities moved from one area to another under the Group Areas Act of 1950, contributing to a housing shortage. A massive program of township development in African areas began in the 1950s but diminished in the 1970s, also contributing to the housing problem. During the 1980s, "site-and-service" schemes emerged to provide land for poorer, usually black people around the cities, but the housing crisis remained severe. In the 1990s, housing policy emphasized the joint roles of the public and private sectors; the government launched an ambitious program of capital subsidies and loan guarantees in an effort to upgrade housing conditions and assist all citizens to acquire title to some form of shelter.

The most important features of social conditions are the high level of unemployment and the wide disparities in wages, both of which redound to the disadvantage of black South Africans. In the mid-1990s more than one-quarter of the African population was unemployed, and those who were employed were generally in the lowest-paying and least prestigious positions. This pattern partially reflects the composition of South Africa's population, with its many migrants to industrial and urban areas. Substantial wage advances in the mining and industrial sectors of

the economy since the 1970s have not been shared by the nonunionized and the unemployed. However, access to government employment, the professions, and business has grown rapidly for Africans, Indians, and those of mixed race, and there are signs of significant change in the distribution of employment in South Africa in the 1990s as more midlevel positions are held by nonwhites.

(M.J.L./D.F.G./A.S.Ma.)

Cultural life. Eleven languages (Afrikaans, English, Ndebele, North Sotho, South Sotho, Swazi [Swati], Tsonga, Tswana, Venda, Xhosa, and Zulu) hold official status under the 1993 constitution, and a further 11 (Arabic, German, Greek, Gujarātī, Hebrew, Hindī, Portuguese, Sanskrit, Tamil, Telugu, and Urdū) are to be promoted and developed. All South African languages are spoken to varying degrees in different regions; there are some areas where most residents speak neither Afrikaans nor English, but those two languages allow communication in most parts of the country. Early school education is available in all the official languages. English appears to predominate to an increasing extent in official, educational, and formal business spheres.

The arts. The many languages spoken by South Africa's people reflect the country's cultural diversity. The Africans have adopted a host of Western ways, but a core of African cultural traditions of language, music, and dance retains its vitality and has contributed to a distinctly South African fusion of cultural forms. Some of the better-known features of that fusion appear in the music of jazz, religious, and popular groups such as the African Jazz Pioneers, Ladysmith Black Mambazo, and others. The various African societies have rich oral traditions, including narrative, poetic, historical, and epic forms. These oral traditions, while they remain strong in their own right, especially in rural parts of the country, have exerted a major influence on the written literatures of the African cultures, which also have been influenced by literary traditions of other parts of Africa, of the Caribbean and the Americas, and of Europe.

Oral traditions

Such writers as Guybon Sinxo (Xhosa), B.W. Vilakazi (Zulu), Oliver Kgadime Matsepe (North Sotho), and Thomas Mofolo (South Sotho) have been more heavily influenced in their written work by the oral traditions of their cultures than by European forms. Works composed in the indigenous languages have been largely ignored or dismissed as works written for the schools or as works with limited audiences, yet these novels and poems have for years been a primary means of expression for African intellectuals. No matter the size and composition of the audiences, the works are a significant part of the intellectual history of South Africa.

During the 1970s there emerged in the arts powerful themes of national and multiracial, multilingual cultural patterns, as writers and artists from all backgrounds concentrated on exploring and portraying the turmoil affecting South African society. Reaction to apartheid engendered a sense of black culture and history that both anticipated and drew upon Negritude as it was manifested in West African, Caribbean, and American movements. The themes of Black Consciousness evident in the poetry and prose of urban writers such as Mothobi Mutloatse and Miriam Tlali and published in such periodicals as *Staffrider* are derived from the literary and oral traditions of African languages in South Africa and in literature by Africans in European languages.

For many decades works with strong political themes or explicit sexual scenes were banned. A large segment of Afrikaner writers became alienated from the government; authors such as Breyten Breytenbach, Dan Roodt, André Brink, and Étienne Leroux had their works banned.

The authors Adam Small and Alex La Guma have written vividly in Afrikaans and English, respectively, of the effects of racial discrimination and of the complex and frequently violent nature of life in South Africa. Many black and white writers addressing these and other themes have received international recognition. Newly recognized writers such as J.M. Coetzee, Sipho Sepamla, and Mongane Wally Serote have joined such established figures as Es'kia (Ezekiel) Mphahlele, Alan Paton, Brink, and Leroux in bringing South African literary life to the wider world. In 1991 Nadine Gordimer, a short-story writer and novelist, was awarded the Nobel Prize for Literature.

Literature

South African playwrights responded to the new cultural and political milieu with such innovations as multilingual plays. Support for the newer indigenous theatre came from independent and nonracial theatrical organizations, such as the Market Theatre in Johannesburg. Plays by Athol Fugard, Mbongeni Ngema, Fatima Dike, and Pieter-Dirk Uys have been performed worldwide.

The major institutional support for culture during the apartheid years came from the provincial councils for the performing arts. These councils helped fund plays, operas, ballets, symphonies, and other cultural events. The postapartheid period brought change in such institutional matters, though under the 1993 constitution the new provinces retained control over cultural affairs.

The press and broadcasting. The press in contemporary South Africa, which has a long tradition of free expression, found itself under increasing political and legal constraints from the 1960s onward. Historically, the strongest elements of the press have been distinct English- and Afrikaans-language publishers. African readership has expanded greatly, though some papers aimed at that market, such as *The World,* were banned during previous decades. Individual journalists were banned, detained, and threatened. During the 1980s a new independent press emerged, represented by newspapers such as *New Nation* and *Weekly Mail. Vrye Weekblad,* the first Afrikaans-language antiapartheid newspaper, closed in 1994. With South Africa's reemergence in the world economy, foreign media interests began to take a greater interest in the local market; the largest daily newspaper group in the country was taken over by an international concern.

Television—introduced in the mid-1970s—and radio constitute important forces in South African society. Until the lifting of emergency media restrictions in February 1990, the government tightly controlled both mediums, which air programs for the different linguistic and cultural groups in the country. The government used television and radio to communicate its own views and to counter cultural tendencies that it perceived to be threatening to the implementation of apartheid. Most electronic media remain publicly owned, but the pattern of management and public participation in their control changed decisively after 1994 as the previously white- and male-dominated management changed to a more representative mix under the new government.

(R.V./D.F.G./A.S.Ma.)

For statistical data on the land and people of South Africa, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The prehistory and history of South Africa span nearly the entire known existence of humans and their ancestors—some three million years or more—and include the wandering of small bands of hominids through the savanna, the inception of herding and farming as ways of life, and the construction of large urban centres. Through this diversity of human experience, several trends can be identified: technological and economic change, shifting systems of belief, and, in the earlier phases of humanity, the interplay between physical evolution and learned behaviour, or culture. Over much of this human career, South Africa's past is also the past of a far wider area, and it is only in the past few centuries that this southernmost country of Africa has had a history of its own.

Prehistory. The earliest creatures that can be identified as direct ancestors of modern humans are classified as australopithecines (literally, "southern apes"), of which the first specimen to be described (in 1925) was the skull of a child from a quarry site at Taung (now in Northern Cape province). Subsequently, more australopithecine bones have been found preserved in limestone caves in the Transvaal region, where they had originally been deposited, up to three million years ago, by predators and scavengers. Australopithecines walked upright, fashioned simple tools from stone and bone, and lived by gathering plant foods and scavenging for meat.

Australopithecine sites

In common with that of other parts of the world, South Africa's prehistory has been divided into a series of phases based on broad patterns of technology. The primary distinction is between a reliance on chipped and flaked stone implements (the Stone Age) and the ability to work iron (the Iron Age). The Stone Age, which in itself spans almost all human history, is further divided into early, middle, and late stages. The simple stone tools found with australopithecine fossil bones fall into the earliest part of the Early Stone Age.

The Early Stone Age. Early humans were evolving both in their physical form, gaining greater dexterity and mental ability, and in their cultural adaptations to the world. In recognition of the physical changes that were taking place, the fossil bones are divided into different genera and species. The genus *Homo* developed from and evolved alongside the genus *Australopithecus* before superseding the more archaic form. Early species of *Homo* also made stone tools that belong to the Early Stone Age; most Early Stone Age sites in South Africa were probably left by people classified as *Homo erectus*.

Simply modified stones, hand axes, scraping tools, and other bifacial artifacts were used for a wide variety of purposes, including butchering animal carcasses, scraping hides, and digging for plant foods. Most Early Stone Age archaeological sites in South Africa are the remains of open camps, often by the sides of rivers and lakes, although people also lived in rock shelters, such as Montagu Cave in the Cape.

The Early Stone Age was a period of very slow change: for more than a million years and over a wide geographic area, there were only the slightest differences in the forms of stone tools. However, the slow alterations in physical appearance that took place over the same time period are sufficient for physical anthropologists to recognize new species in the genus *Homo*. By about 500,000 years ago, some people were sufficiently modern in appearance to be considered an archaic form of *Homo sapiens:* important specimens belonging to this physical type have been found at Hopefield in Western Cape and at the Cave of Hearths in Mpumalanga.

The Middle Stone Age. Sometime after 200,000 years ago this long episode of slow cultural and physical evolution gave way to a period of more rapid change. Hand axes and big bifacial stone tools were no longer made, and people began to use stone flakes and blades to fashion scrapers, spear points, and parts for hafted, composite implements—a technological stage that has become known as the Middle Stone Age.

Change to flaked tools

There are numerous Middle Stone Age sites in South Africa. People continued to live in open camps, while rock overhangs also were used for shelter. Day-to-day debris has survived to provide some evidence of early ways of life, although plant foods, which must have been important, are rarely preserved. Middle Stone Age bands hunted medium-size and large prey such as antelope and zebra, although they tended to avoid the largest and most dangerous animals such as the elephant and rhinoceros. Sometimes they collected tortoises and ostrich eggs in large quantities, as well as seabirds and marine mammals that could be found along the shore. The rich archaeological deposits of Klasies River Mouth Cave, on the Cape coast west of Port Elizabeth, preserve the earliest evidence in the world for the use of shellfish as a food source.

Klasies River Mouth Cave also is important for the evidence it provides for the emergence of anatomically modern humans. Some of the human skeletons from the lower levels of this site, possibly as old as 115,000 years, are modern in form. Equally early fossils have been excavated at Border Cave, in the mountainous region between KwaZulu/Natal and Swaziland. This archaeological evidence is consistent with the results of research by geneticists, some of whom believe that there was a single ancestral population of modern humans living in Africa about 200,000 years ago.

The Late Stone Age. About 40,000 years ago people again began to change their basic techniques of toolmaking. Small, finely worked stone implements known as microliths began to become common, and the heavier scrapers and points of the Middle Stone Age less frequent. Archaeologists refer to this new technological stage as the Late Stone Age. The new ways of working stone again reflect an accelerating pace of change. In contrast to the almost static millennia of the Early Stone Age and the slow cultural changes of the Middle Stone Age, Late Stone Age people were far more responsive to changes in their environment. The numerous collections of stone tools from South African archaeological sites dating to the past 40,000 years show, in consequence, a great degree of variation through time and across the subcontinent.

Like their predecessors, Late Stone Age people relied heavily on plant foods, the remains of which have been well preserved at sites in the Cape region. Animals were trapped and hunted with spears and arrows on which were mounted fine stone blades. In many areas people moved with the seasons, following game into higher lands in the spring and early summer months, when new flushes of plant foods could also be found. When available, rock overhangs were occupied; otherwise, windbreaks were used for shelter. Coastal resources were important, resulting in numerous shell middens scattered along the full length of the South African coastline. Shellfish, crayfish, seals, and seabirds were collected or caught along the shore, and fish were caught on lines, with spears, in traps, and possibly with nets.

Belief systems

Late Stone Age communities are the first to have left evidence for complex systems of belief, probably because sophisticated symbolic abilities are uniquely part of the anatomically modern condition—earlier forms of humans probably did not think in this way. There are numerous engravings on rock surfaces, mostly on the interior plateau, and paintings on the walls of rock shelters in the mountainous regions of South Africa, such as the Drakensberg and Cedarberg ranges. Dating is difficult, but it is clear that the art spans at least the past 25,000 years. South African rock art was originally seen either as the work of exotic foreigners such as Minoans or Phoenicians or as the product of primitive minds. Now it is widely accepted that the paintings were closely associated with the work of medicine people: shamans who were involved in the well-being of the band and often worked in a state of trance. Specific representations include trance dances, metaphors for trance such as death and flight, rainmaking, and control of the movement of antelope herds.

Pastoralism and early agriculture. About 2,000 years ago, new ways of living came to South Africa. From their earliest years, human communities had lived by gathering plant foods and by hunting, trapping, and scavenging for meat. However, people began to make use of domesticated animals and plants. In the east, where rainfall is adequate, crops could be grown and cattle, sheep, and goats herded near permanent villages and towns. In the west, where the climate is arid or where rain falls at the wrong times of the year for African cultivated plants, domestic livestock were kept by nomadic pastoralists, who moved over wide territories with their flocks and herds.

Early nomadic pastoralism

The origin of nomadic pastoralism in South Africa is still obscure. Linguistic evidence points to northern Botswana as a centre of origin, and this is supported by sheep bones, found in the same archaeological levels as pottery, that have been dated to about 150 BC from Bambata Cave in southwestern Zimbabwe. It is still unclear whether new communities moved into South Africa with their flocks and herds or whether established hunter-gatherer bands took up completely new ways of living. However, the results of archaeological excavations have shown that by the first few centuries AD sheep were being herded fairly extensively in Eastern and Western Cape provinces and probably in Northern Cape as well.

Surviving traces of sites where herders lived tend to be elusive. One of the best-preserved camps is at Kasteelberg, on the southwest coast near St. Helena Bay, where pastoralists kept sheep, hunted seals and other wild animals, and gathered shellfish, repeatedly returning to the same site for some 1,500 years. Such communities were directly ancestral to the Khoikhoi herders who encountered European settlers at the Cape of Good Hope in the mid-17th century.

The archaeological traces of farmers in the eastern regions of South Africa are more substantial. The earliest sites date to the 3rd century AD, although it is probable that farmers were already well established by this time. Scatters of potsherds with distinctive incised decoration mark early village locations in Mpumalanga and parts of KwaZulu/Natal.

The Iron Age. These first farmers had knowledge of ironworking, and their archaeological sites are grouped together as the Iron Age. The Early Iron Age represents the arrival in South Africa of new groups of people, having strong connections with East Africa and directly ancestral to the Bantu-speaking communities who form the majority of South Africa's population today.

Early Iron Age farmers grew crops, cutting back the vegetation with iron hoes and axes, and herded cattle and sheep. They also relied heavily on gathering wild plant foods, some hunting, and collecting shellfish if they lived near enough to the coast. Where conditions for agriculture were favourable (such as in the Tugela [Thukela] River valley in KwaZulu/Natal), villages grew to house several hundred people. There was probably some trade between different groups of farmers—evidence for specialization in salt making has been found in Mpumalanga—and with the hunter-gatherer bands that continued to occupy most parts of South Africa. Finely made, life-size ceramic heads from Lydenburg in Mpumalanga, dated to the 7th century AD, are a shadow of the complex systems of belief that have largely been lost to history.

Early Iron Age villages were built in low-lying areas, such as river valleys and the coastal plain, where forests and savannas allowed slash-and-burn agriculture. However, from the 11th century, farming communities began to settle the higher-lying grasslands beneath the Drakensberg and on the interior plateau. In many areas they started making different forms of pottery as well as villages built of stone. It is probable that these and other changes in patterns of behaviour reflect the increasing importance of cattle in both economic and social life. By convention, this later phase of precolonial farming is known as the Late Iron Age.

Other changes came in the north. Arab traders had begun to establish small settlements on the Tanzanian and Mozambican coasts in their search for ivory, animal skins, and other exotica. The trade beads they offered in return began to reach villages in the interior, the first indications that the more complex economic and social structures associated with long-distance trade were developing. The arid Limpopo River valley, eschewed by the earliest farmers, became a focal point of settlement as a natural trade route. Sites such as Pont Drift (about AD 800 to 1100) and Schroda (dated to the 9th century) show that their occupants were rich in both livestock and trade beads.

Mapungubwe

This was the setting in which Mapungubwe developed as South Africa's first urban centre. Starting as a large village like Schroda and Pont Drift, Mapungubwe rapidly developed into a town of perhaps 10,000 people. Differences in status were marked out clearly: the elite lived, and were buried, on the top of the stark sandstone hill that is at the centre of the town, while ordinary people lived in the valley below. The hilltop graves contained lavish burial goods, including a carefully crafted gold rhinoceros, and excavations have provided evidence for specialized crafts such as bone and ivory working, all suggestive of social and economic differentiation. Mapungubwe was abandoned sometime in the 12th century, after having been occupied for perhaps 200 years. It is likely that both the town's founding and its decline were closely related to trade—to the possibilities that the connection with the coast along the Limpopo valley offered for the accumulation of wealth and to the capture of this trade by Great Zimbabwe, farther to the north.

Europeans had long envied the riches of the Arab world, and in 1488 the first Portuguese ships rounded the Cape of Good Hope, pursuing a share of the lucrative Arab trade with the East. Over the following century, numerous vessels made their way off the South African coast, but the only direct contacts came with the bands of shipwreck survivors who either set up camp in the hope of rescue or tried to make their way northward to Portuguese settlements in Mozambique. From the early 17th century, the Portuguese control of the Cape sea route was challenged by both the British and the Dutch. In 1615 the British founded a short-lived settlement at Table Bay, and in 1652 the Dutch East India Company set up a small garrison under the slopes of Table Mountain, charged with the task of provisioning the Dutch fleets.

Settlement of the Cape Colony. The Dutch East India Company, always mindful of unnecessary expense, did not intend more than a minimal presence in the southernmost part of Africa. However, farming beyond the shores of Table Bay proved necessary, and in 1657 nine men were released from their contracts with the company and granted land along the Liesbeek River. In the same year, the first slaves were brought to the Cape, and, by the end of the century, the stamp of Dutch colonialism in South Africa was clear. Settlers, aided by increasing numbers of slaves, grew wheat, tended vineyards, and grazed their sheep and cattle from the Cape peninsula to the Hottentots Holland Mountains, some 30 miles away. A census of 1707 listed 1,779 settlers, owning 1,107 slaves.

In the initial years of Dutch settlement at the Cape, Khoikhoi pastoralists were keen to trade. However, the garrison's demand for cattle and sheep seemed insatiable, and the Khoikhoi became more wary. The Dutch began to push farther into the interior, seeking livestock in return for tobacco, alcohol, and trinkets. Numerous conflicts followed, and many Khoikhoi communities were decimated by smallpox, particularly in 1713. At the same time, colonial pastoralists, or trekboers, began to move inland beyond the Hottentots Holland Mountains with their own herds. By the end of the 18th century, the Khoikhoi chiefdoms had been largely decimated, their people either dead or reduced to conditions close to serfdom on colonial farms. The San (or Bushmen)—small bands of hunter-gatherers who had hung on to old ways of life in isolated areas—fared no better. Pushed back into marginal areas, they were forced to live by cattle raiding, justifying in colonial eyes their systematic eradication. Men were slaughtered and women and children taken into servitude.

Expansion into the interior

Trekboers were in constant search of new pasturage, and they and their families spread northeast as well as north, moving toward the grasslands long occupied by Late Iron Age farmers. For many generations these communities had lived in settlements concentrated along the low ridges of the interior plateau. Population estimates are difficult, but some of these larger villages must certainly have housed several hundred people. Cattle were kraaled (penned) in elaborately built stone enclosures, the ruins of which survive today across a large part of Free State and in the higher areas of the Transvaal region. Extensive networks of exchange brought iron for hoes and spears from specialized manufacturing centres in the Mpumalanga Lowveld and the deep river gorges of KwaZulu/Natal.

Thus, by the closing decades of the 18th century, South Africa fell into two broad regions. The west—including the winter rainfall region around the Cape of Good Hope, the coastal hinterland northward toward the Namibian border, and the dry lands of the interior—was dominated by the advancing frontier of colonial settlement. Trekboers were taking increasingly more land from the Khoikhoi and from remnant hunter-gatherer communities, who were killed, were forced into marginal areas, or became labourers tied to the farms of their new overlords. In the east—where summer rainfall and good grazing made mixed farming economies possible—Late Iron Age farmers had long been established in both coastal and valley lowlands and on the Highveld of the interior.

Growth of Cape Town

Meanwhile, Cape Town had been developing as South Africa's second urban centre, although it was many years before it reached the size of Mapungubwe some five centuries earlier. The administration of town and colony was in the hands of a governor and council, and the economy was in principle directed by the economic interests of the Dutch East India Company; in practice corruption and illegal trading were the order of the day. Both town and colony depended for their existence on the slaves, who by now outnumbered their owners. (M.Hl.)

Growth of the colonial economy. During the century from about 1770 to about 1870, the region now known as South Africa was integrated more fully into the world capitalist economy. The decisive moment was the seizure of the Cape Colony by Britain in 1806 as a strategic base (securing the developing empire in India), market, source of raw materials, and outlet for surplus people. Britain possessed world hegemony between the 1810s and '60s, and its shadow loomed over southern Africa.

By the 1860s the Cape Colony had spawned the subcolonies of Natal, the Orange Free State, and the Transvaal. Whites settled to the edges of the Kalahari desert in the west, the Drakensberg and Natal coast in the east, and the tsetse-fly- and mosquito-ridden Lowveld along the Limpopo River valley in the northeast. Africans were dispossessed of much of their land, and many of them were forced to work for the settlers. The settler population increased from about 20,000 in the 1780s to about 300,000 in the late 1860s. It is impossible to estimate the African population accurately, but it is likely to have been between 2,000,000 and 4,000,000.

Evolution of the African states

After the 1760s African societies were increasingly affected by ivory and slave traders operating from Delagoa Bay, Inhambane, and the lower Zambezi (in modern Mozambique) as well as by traders and raiders based in the Cape to the south. In response to these invasions, surviving African farming peoples evolved a number of sister states different in structure, scale, and military capacity from anything that had gone before. The most successful were the Pedi and Swazi in the eastern Highveld, the Zulu south of the Pongola River, the Sotho in the Caledon valley, the Gaza along the lower Limpopo, and the Ndebele in southwestern Zimbabwe. Possessing agriculture, these African societies proved resilient, unlike the hunting societies in the Cape, the Americas, and Australia. But unlike, for instance, states in China, they were unable to repel the invaders altogether, possibly because food surpluses had been insufficient to support a military class able to force unification early enough. Nor, even at their height in the 1860s, were the African states able to unite. They were able to protect their peoples from proletarianization and impoverishment for only a brief period. Accentuated competition among the European nations and the discovery of gold and diamonds in the 1860s led to a new imperialist offensive: between the 1870s and '90s the African states were destroyed by Europeans exploiting African divisions and using new breech-loading rifles and the first machine guns.

After the 1770s the Dutch settlers and African farming peoples collided. Dutch trekboers advanced across the semidesert Karoo of the central Cape and met agricultural peoples along a line running from the lower Vaal and middle Orange river valleys to the sea around the Gamtoos River (west of Port Elizabeth). West of this line the nonagricultural Khoi groups had already been badly disrupted by European invaders. The Boers were weakly controlled by the Dutch East India Company, a mercantilist organization that monopolized a sparse trade and was unenthusiastic about colonization. Rebellions against the company—as, for instance, by the burghers of Swellendam and of Graaff-Reinet in 1795—became frequent. In the 1780s armed clashes over land and cattle began between the Boers and Xhosa groups such as the Gqunukhwebe and Ndlambe in the Zuurveld between the Sundays and Fish rivers. The Boers developed a commando system in these continuous forays.

In contrast to this eastern frontier, the longer-settled areas of the western Cape had evolved a less fluid agricultural economy of grape and wheat farms run by imported slave labour. Slaves were treated harshly, and punishments for assaulting whites were brutal—for instance, death by impalement. Escaped slaves formed Maroons—small, self-sufficient communities—or fled into the interior. Because slave birth rates were low and settler numbers were increasing, the Dutch in the 1780s stepped up the enserfment of surviving Khoi (pejoratively called "Hottentots") to help run the farms. Many Khoi in the east joined Xhosa groups in a major counteroffensive against colonialism in 1799–1801, and there were slave rebellions in 1808 and 1825.

The Dutch refusal to grant citizenship rights (*e.g.*, access to land) to "Coloured" offspring of unions between whites and Khoi or slaves produced aggrieved "Bastard" groups that were Christian, spoke Dutch, and had an excellent knowledge of horses and firearms. Many fled north toward and over the Orange River in search of land and trading opportunities. After merging with independent Khoi groups, such as the Kora, they formed commando states under warlords, the more successful of whom came from the Bloem, Kok, Barends, and Afrikaner families. By the 1790s they were trading with and raiding local African communities such as the Rolong, Tlhaping, Hurutshe, and, farthest north, Ngwaketse. These groups coalesced into larger aggregations of people, who competed with each other to control trade routes going south to the Cape and east to Mozambique. By about 1810 Moleabangwe of the Tlhaping and Makaba of the Ngwaketse, for example, were already powerful rulers.

European trade on the eastern coast

Along the southeast coast the Portuguese and some British and French were trading beads, brass, cloth, alcohol, and firearms in return for ivory, slaves, cattle, gold, and minor items such as wax and skins. During the late 18th century, high volumes of ivory were exported from Delagoa Bay annually, and slaves were taken from the Komati and Maputo river regions and sent to the Mascarene Islands in the Indian Ocean and to Brazil. There they worked sugarcane, coffee, and other plantations that met the needs of Europe's rising population. By 1800 trade routes linked Delagoa Bay northwest to the Soutpansberg (where copper was obtained from the Venda), west to the Tlhaping and Hurutshe, and south to the Zuurveld—thereby linking the Cape and coastal trade routes with the central interior.

This European trade was a primary cause of structural transformation within societies inland of Delagoa Bay. Warlords reorganized military institutions to hunt elephants and slaves. Profits from trade enhanced patronage capabilities, attracted followers, and raised military potential and, in turn, the capacity to dominate land, people, and cattle. Near the bay, Tembe and Maputo were already powerful states by the 1790s. To the west emerged the Maroteng of Thulare, the Dlamini of Ndvungunye, and the Hlubi of Bhungane. Between the Pongola and Tugela rivers evolved the Mthethwa of Dingiswayo south of Lake St. Lucia, the Ndwandwe of Zwide, the Qwabe of Phakatwayo, the Chunu of Macingwane, and, south of the Tugela, the Cele and Thuli. These groups competed to dominate trade and were the more militarized the closer they were to the Portuguese base.

Accentuated European impact, *c.* 1810–35. *British occupation of the Cape.* In 1795 the British responded to France's overrunning the Dutch Republic by occupying the Cape. After returning it at the Treaty of Amiens in 1802, they reannexed the colony in 1806 after the beginning of the Napoleonic Wars. Prior to the opening of the Suez Canal in 1869, the Cape became a vital base for Britain—the Cape economy was meshed with Britain's. London Missionary Society and Methodist missionaries, as well as hunters and traders, ranged into Transorangia (the territory in the interior to the north and south of the Orange River) and the Transkei, mapping and exploring. Until 1825 Cape wines were given preferential access to the British market. Merino sheep were introduced, and serious sheep farming was begun to supply wool to British textile mills.

The infrastructure of a new type of colony was established. English replaced Dutch as the language of administration; in 1825 the Dutch rix-dollar was replaced by sterling; newspapers opened in Cape Town after 1824; British colonial governors were brought out; and in 1825 an advisory council for the governor was established, which was upgraded into a legislative council in 1834 with a few "unofficial" settler representatives. In 1813 most of the British East India Company's privileges were abolished (the company's last monopoly on the Chinese trade ended in 1833), after which the number of merchants increased and shipping trade expanded. After 1813 the Dutch loan farm system—whereby white colonists paid a small annual fee to the government but did not acquire ownership of

the land—was gradually replaced with a virtual freehold system of landownership.

British drive toward the Fish River

In 1820 a large group of British settlers arrived. Together with a high white birth rate and wasteful land usage, this produced an accentuated land hunger. To secure colonial hegemony and to alleviate the settler land shortage, the British applied massive military intervention against Africans on the eastern frontier. Until the 1840s the British did not envisage the colony expanding to include African ("Kaffir") citizens: the latter were to be expelled across the Fish River, the unilaterally proclaimed eastern border of the colony.

The first step in this process was the onslaught of the British army against the Zuurveld Gqunukhwebe and Ndlambe in 1811–12. In the aftermath of this war, Graham's Town (now Grahamstown) was established as the military pivot of a line of forts. The removal of about 20,000 people eastward across the Fish aroused tensions that British governors exploited by divide-and-rule strategies. This exacerbated tensions between the Ngqika and the Ndlambe and Gcaleka, leading to the Battle of Amalinda in 1818. The Thembu and Mpondo farther east were established as British allies against all three of these peoples. An Ndlambe attack on Graham's Town in 1819 provided the pretext for the annexation of the next swath of African territory, to the Keiskamma River. The 1820 settlers took the African lands. Khoi soldiers were settled along the Kat River in 1829, and Rharhabe groups (*e.g.*, Maqoma's) were repeatedly harried from their lands in the early 1830s. When the Rharhabe despairingly counterattacked in December 1834, Governor Benjamin D'Urban ordered the colossal invasion of 1835; thousands of Rharhabe were killed. In April 1835 the British crossed the Great Kei River and ravaged Gcaleka territory; on May 12 they murdered the Gcaleka chief, Hintsa. Only the intervention of the British colonial secretary, Lord Glenelg (to the fury of the settler expansionists), halted the seizure of all African land to the Great Kei. D'Urban's pioneering attempt to rule conquered Africans with white magistrates and soldiers was overturned by Glenelg; instead, for a time, Africans east of the Keiskamma retained their autonomy and dealt with the colony through diplomatic agents.

A chronic problem for the British was how to procure labour to develop the new settler farms and to build the towns. Britain abolished its slave trade in 1807 and pressured other nations to do the same. Some slaves continued to be imported into the Cape after 1807 (for example, "prize negroes"—slaves seized by the Royal Navy and reenslaved in the Cape), but they were not enough. An 1809 ban on Africans crossing the border aggravated the labour shortage, and the British, like the Dutch, were forced to enserf the Khoi (by the Caledon and Cradock codes of 1809 and 1812).

Anglo-Boer commandos illegally captured San women and children (exterminating many of the men), as well as Africans from across the eastern frontier. More serfs (called "apprentices") were captured by the Griqua raiding states led by Andries Waterboer, Adam Kok, and Barend Barends along the Vaal, Molopo, and Caledon rivers from among the Taung, Hurutshe, Rolong, Kwena, and Fokeng peoples. The prisoners, known as Mantatees, were imported mainly into the eastern Cape in exchange for firearms, gunpowder, and horses; they were set to work on the farms. One Griqua raid on an unidentified people at Dithakong in 1823 was accompanied by the British government agent John Melvill and the Kuruman missionary Robert Moffat. White farmers also raided for labour north of the Orange River.

Changes in labour policy

The need to align the Cape with the growing imperial antislavery ethos and to facilitate labour distribution produced an overhaul of labour policy in 1828. Ordinance 50 freed the Khoi to choose their employers but brought little tangible change to their positions; innumerable disincentives existed if they preferred not to work. Ordinance 49 permitted black labourers from east of the Keiskamma to come into the colony for work under control of contracts and passes issued by soldiers and missionaries. At first this permission was insufficient, and in August 1828 Anglo-Boer armies (supported by Khoi, Thembu, Gcaleka, and Mpondo auxiliaries) attacked the Ngwane east of the Great Kei at Mbolompo, returning with prisoners. More spectacularly, in its operations of 1835, the British army and its Khoi and African collaborators seized thousands of Rharhabe and Gcaleka women and children. Known as "Fingo," they were dispatched as labourers throughout the Cape, so that by 1837 the Cape Colony's labour supply was assured, though only temporarily. Ironically, these illegal measures coincided with the formal abolition of slavery in 1834–38. Ex-slaves, Khoi and Fingo, were now controlled by the Masters and Servants Ordinance of 1841. This imposed criminal penalties for breach of contract and desertion of the workplace and increased the legal powers of settler employers.

Historic states of South Africa.

The Delagoa Bay slave trade. Concurrent with these events in the Cape, the slave trade at Delagoa Bay had been expanding since about 1810 as the Brazilian plantations grew. During the late 1820s, slave exports from the Delagoa Bay area reached several thousand a year, in anticipation of what proved to be an ineffective attempt to abolish the Brazilian trade in 1830. After a dip in the early 1830s, the Bay slave trade reached a peak in the late 1840s.

The impact on hinterland societies was increasingly profound. Makhasane's Maputo and other groups became surrogate slavers and joined the Portuguese soldiers in inland raiding. Along the Limpopo and Vaal river networks Bay slavers competed with Griqua slavers supplying the Cape. The Gaza and Jele near St. Lucia moved north to slave; the Jele, or Ngoni, later resituated themselves west of Lake Nyasa, liaising with Arab slavers. Slavers burned crops, and famines were common. Many groups, including the Ngwane, Ndebele, and some Hlubi, fled westward into the Highveld mountains during the 1810s and '20s. The Ngwane were attacked in sequence by Bay, Griqua, and British slavers. The Patsa (Kololo) of Sebetwane, on the other hand, moved east out of Transorangia, ran into Bay slavers, migrated west into Botswana, where in 1826 they were attacked by an alliance of Ngwaketse and white mercenaries, and ended in Zambia in the 1850s exporting slaves to the Arabs and Portuguese. These migrations all produced further destabilizations.

Emergence of the eastern states. By the 1820s four main defensive state clusters had emerged between the Soutpansberg and the Drakensberg: Sekwati's Pedi in the Steelpoort valley, Sobhuza's Dlamini in the eastern Transvaal, the Mokoteli of Moshoeshoe (Mshweshwe) in the Caledon River region, and Shaka's Zulu south of the Black Mfolozi River. The Pedi received refugees from the Limpopo and coastal plains. Moshoeshoe's people absorbed refugees from throughout eastern Transorangia;

they became formidable raiders and by the mid-1830s were able to defeat the Griqua and Korana raiders. Between about 1812 and 1825 Shaka welded the Chunu, Mthethwa, Qwabe, Mkhize, Cele, and other groups into a militarized state with fortified settlements called *amakhanda*. Zulu *amabutho* (regiments) defended against raiders, provided protection for refugees, and, the evidence suggests, began to trade in ivory and slaves themselves.

From 1824 the Zulu faced competition from Cape colonists who came to Port Natal (renamed Durban in 1835) and organized mercenary armies. Although on a smaller scale, these were comparable to the Portuguese *prazero* armies along the Zambezi and to the warlord state set up by the Portuguese trader João Albasini in the Soutpansberg in the 1840s. During the 1820s European raiders joined Zulu *amabutho* in raids north of the Black Mfolozi River and also operated south of the Mzimkhulu River—where slaves were being exported on French ships in 1825. In 1828 Francis Farewell's raiders, in alliance with Zulu groups, seized women and children in the same area. Trade conflicts helped split the Zulu elite into rival factions, a split that led to Shaka's assassination in 1828. The succession of Shaka's half brother, Dingane, was accompanied by civil wars and by increasing interference in the Bay trading alliances. In 1833 the Portuguese governor at Lourenço Marques (now Maputo), Ribeiro, was killed in one of these wars. White warlords in Natal such as Henry Flynn, Nathaniel Isaacs, and John Cane continued to slave and hunt elephants. By the mid-1830s consortia of Cape merchants were planning the formal colonization of Natal with its superb agricultural soils and temperate climate. The British left the less desirable Delagoa Bay region to the Portuguese, who traded slaves out of Lourenço Marques for another half-century.

Raids of the 1820s

The expansion of white colonialism, *c.* 1835–1870. *The Great Trek.* After about 1834 a previous trickle of Boer migrants north of the Orange River suddenly became an organized flood. This later became known as the "Great Trek." The common view that this was a bid to escape the policies of the British—for instance, the freeing of the slaves—is difficult to sustain, as most of the ex-slave owners did not migrate (most trekkers came from the poorer east Cape), and in 1835 the labour shortage had been alleviated. The trek was the explosive culmination of a long sequence of colonial labour raids, grazing probes, land seizures, punishment commandos, and commercial expansions. Whites possessed weaponry that was always technologically one step ahead of that of the Africans. They also had the instructive examples of how small groups of raiders in Natal and Transorangia had wrought havoc over large areas and how the British army had induced terror in Africans. The trekkers were not backward feudalists escaping the modern world, as some historians have maintained; they were energized people extending their frontier. The trek was as inevitable a development as the North American colonists' push across the Alleghenies in the 1760s.

Several thousand Boers migrated with their families, livestock, retainers, wagons, and firearms into a region already destabilized and partially depopulated by Griqua and coastal raiders. Only when they came up against Mzilikazi's Ndebele (who in the early 1830s had moved from the southeastern to the western Transvaal) were they confronted, as at Vegkop in 1836. However, the Boers—in alliance with Rolong, Taung, and Griqua allies—crushed the Ndebele during 1837, taking their land and many cattle, women, and children. The Ndebele fled north, resettling around the Matopo and Malungwane hills in Zimbabwe.

By the early 1840s the trekkers had penetrated much of the Transvaal. A grouping of commando states emerged based on Potchefstroom, Pretoria, and, from 1845, Ohrigstad-Lydenburg in the eastern Transvaal. Andries Hendrik Potgieter, Andries Pretorius, Jan Mocke, and others competed for followers, attacked weaker African chiefdoms, hunted elephants and slaves, and forged trade links with the Portuguese. The development of farms was slow and inevitably depended on forced labour, as had been the case in the Cape prior to the 1830s. For a quarter of a century until the 1860s, moreover, the Pedi and Swazi in the east and even Kwena and Hurutshe groups in the west were strong enough to curb Boer expansionism.

Other Boers turned east into Natal and allied themselves with the resident British settlers. There was an inescapable confrontation between this coalition and Dingane's Zulu. The Zulu, though like the Ndebele scoring initial successes, were overpowered at Blood (Ncome) River in 1838 and the AmaQonqo Hills in 1840. The Boers, aided by Zulu civil war, annexed land to the Black Mfolozi River and set up Mpande as a puppet over the much-reduced Zulu kingdom. Colonial military strategies perfected along the Fish and Great Kei rivers in the eastern Cape were transplanted to the Tugela and Pongola. The whites began to carve out farms in Natal as they had done along the eastern frontier. Further slave and cattle raids on Ncaphayi's Bhaca south of the Mzimkhulu provided the pretext for British annexation of Natal in 1843. The Zulu were returned land between the Mfolozi and Tugela and for the time being left independent. Mpande (reigned 1840–72) was a formidable ruler and further built up Zulu military capacity, which his son, Cetshwayo, used effectively against the British invaders at Isandlwana in 1879.

The move into Natal

The British in Natal. The coming of thousands of British settlers to Natal in the 1840s and '50s meant that for the first time black Africans and white settlers lived together (however uneasily) on the same land. In 1845 a diplomatic agent (later secretary of native affairs) was appointed: Theophilus Shepstone—a prototype of later chief native commissioners. Reserves for blacks were set aside (Harding Commission, 1852), and missionaries and pliant chiefs were introduced to persuade them to work. After 1849 Africans were subjected to a hut tax that was designed to raise revenue and force them into labour; *isibhalo* (forced) labour was used for road building; and Africans on state land and white farms were made to pay rents. To meet these burdens, the more resilient African cultivators—or squatters—grew surplus crops to sell to the growing towns of Pietermaritzburg and Durban.

The British were reluctant, though, to annex the Transorangian interior. No strategic interests were involved. Boer trade links with Delagoa Bay posed little threat, as Portugal was a virtual client state of Britain. The tasks of eroding African resistance and developing the land were left to the Boers. The policy was muddled and never clearly enunciated. Financial constraints operated. A half-hearted attempt was made to protect Britain's long-standing Griqua client states. Sir George Napier in 1843 and Henry Warden in 1849 attempted to arbitrate a border between Moshoeshoe's Sotho state and the Boers west of the Caledon River. After further war with the Rharhabe on the eastern frontier in 1846, the aggressive governor, Colonel Harry Smith, in 1847–48 finally annexed the regions between the Fish and the Great Kei rivers (establishing British Kaffraria) and between the Orange and Vaal (Orange River Sovereignty). These moves provoked further war with the Xhosa (joined once more by many Khoi) in 1851–53 and a free-for-all in the sovereignty, with unsupported British politicians ineffectively trying to influence events.

Smith's annexation

A striking feature of this period was the capacity of the Sotho people to fend off military conquest by the British and Boers. After defeating and absorbing the rival Tlokwa of Sekonyela in 1853–54, Moshoeshoe became the most powerful African leader south of the Vaal-Pongola. His soldiers utilized firearms and, in the cold Highveld, horses—the keys to political and military survival.

Attempts at Boer consolidation. Faced with these unprofitable conflicts, the British temporarily withdrew, and the Transvaal and Orange Free State Boers were given independence at the Sand River and Bloemfontein conventions of 1852 and 1854, respectively. The access of Africans to guns and gunpowder was prohibited. Both Boer groups wrote constitutions and established Volksraade (parliaments), although their attempts at unification failed. For more than a decade consolidation among the Boers was hampered by civil wars and the struggle with the material environment. Nevertheless, the Orange Free State's economy grew rapidly, and by the 1860s the Boers were exporting significant amounts of wool via Cape ports.

The Cape economy. Capitalist infrastructure came earlier in the Cape because of its older colonialism and its seacoast links to the empire. Banks and insurance and limited-liability companies were founded in the 1840s and '50s. The ending of ceilings on interest rates in 1860 attracted capital, which was loaned against rising property values. A class of prosperous colonial shopkeepers, financiers, traders, and farmers emerged. Cape Town grew to more than 30,000 people in the 1850s. Port Elizabeth, established in 1820, became an important trading centre and harbour. Representative government came in March 1853: the Legislative Assembly had elected members, but an executive was appointed from London; only in 1872 was the executive made responsible to the assembly. Franchise qualifications were relatively low. Some Africans could even vote, but the numbers were too small to have political impact. These nominal rights were reduced later in the century and abolished outright in 1936.

Legal restrictions on land and labour

The colonial attacks of 1811, 1819, 1835, 1846, 1851, and 1858 deprived Africans of most of their land between the Sundays and Great Kei rivers and produced impoverishment and despair. From 1855 British magistrates were imposed in British Kaffraria, and the power of the Xhosa chiefs was destroyed. Following a severe lung sickness epidemic in 1854–56, the Xhosa killed many of the remaining cattle and in 1857–58 were unable to grow many crops (for reasons that are still not entirely clear). The subsequent starvation drove yet more thousands of Africans into the Cape Colony to work. In British Kaffraria, as in Natal, the hut tax was introduced in 1849. African citizens of the Cape Colony (Fingo) were controlled under the Masters and Servants Act of 1856 (an updating of Ordinance 50 of 1828 and the 1841 ordinance), and non-Cape Africans were regulated by the Kaffir Employment Act of 1857 (an updated version of Ordinance 49). In 1865 British Kaffraria was fused with the Cape Colony, and thousands of newly defined Fingo resettled east of the Great Kei, thereby creating Fingoland. The Transkei—the hilly country between the Cape and Natal—became a large African reserve, the still independent parts of which were annexed in the 1880s and '90s. Mpondoland fell in 1894.

White missionaries and their black catechists worked sedulously from the 1820s to undermine African cosmologies and seduced Africans into desiring European manufactures they had previously done well without. The techniques for destroying African cultures and for forcing Africans to work pioneered in the Cape and Natal were exported to the rest of Africa after the 1880s. For a time, nevertheless, there was room for a small class of African peasant farmers (producing for the market and liable for rents and taxes), who used plows and sold surplus grain to the towns in competition with colonial farmers. Difficulty in obtaining capital, as well as legal and political discrimination, drove most of them out of business in the decades following the South African War of 1899–1902.

The Cape economy, narrowly based on wine and wool, was not particularly prosperous. Wool exports, though soaring to 12,000,000 pounds (5,440 metric tons) in 1855, lagged far behind those of Australia and were susceptible to drought and market slumps, as in the early 1860s. Roads were built with African labour, but only a few miles of railway were constructed before 1870. Attempts to broaden the economic base were not at first successful. Guano (droppings of gannets and cormorants used as fertilizer) was exploited on off-coast islands; copper began to be mined in Namaqualand; hunters operating as far as the Zambezi sent out much ivory; and traders, hunters, missionaries, and full-time prospectors surveyed and sampled the rocks. The efforts of the last were rewarded with the discovery of diamonds in the Vaal valley and of gold in the Tati valley in 1866–67 and in the northern and eastern Transvaal in 1871.

Disputes in the north and east. To the north, colonial communities and the African states cooperated and competed with each other, the advantage slowly moving to the colonists. Mswati's Dlamini Swazi and Manukosi's (Soshangane's) Gaza supplied slaves both to the Transvaal Boers and to the Portuguese. Mswati's people overran much of the Lowveld, incorporating many groups (the *emakhandzambile*) and exchanging captured children for firearms and horses with the Transvaal settlers. The death of Manukosi in 1858 led to a Gaza civil war in which the Swazi, the Boers, and the Portuguese all intervened. In 1864, when Mswati controlled the land almost to Lourenço Marques, the Gaza (under the victor, Mzila) migrated northward into the Búzi River area of eastern Zimbabwe.

Swazi-controlled areas

Farther south the Zulu competed with the Swazi and the Boers to dominate the Pongola and Ingwavuma valleys and with the Boers to control the Buffalo (Mziniathi) River area. A Zulu civil war in 1856 (the Battle of Ndondakasuka on the lower Tugela River) elevated Cetshwayo over Mbuyazwe, and he effectively ruled Zululand from the early 1860s. Shepstone interfered not only in Zulu politics but also in an Ndebele succession dispute, attempting to oust Lobengula in favour of a pretender in 1869–72. Marthinus Pretorius, the Transvaal leader, annexed huge areas, at least on paper. To the irritation of settler farmers and plantation owners, few Zulu went south to work in Natal. Instead, a supply of Mozambican indentured labourers (some of them effectively forced) was organized. This evolved in the following decades into a steady flow of migrant labour. However, initially there was not enough labour to satisfy the new sugar plantations, and, from about 1860, indentured labourers from India were brought over to do the work.

Moshoeshoe's Sotho continued their tenacious hold on their lands along the Caledon River and were for a time able to supply the Boers of the Orange Free State with grain and cattle. The Boers, however, coveted the fertile Caledon valley. Moshoeshoe mobilized 10,000 men to defeat them in the war of 1858; but, with the consolidation of the Orange Free State under Johannes Henricus Brand in the early 1860s, the Sotho were defeated in 1865–66 (Treaty of Thaba Bosiu, 1866), and only British annexation of Moshoeshoe's territory in 1868 prevented complete Sotho collapse.

The decline of the African states. As the 1860s came to an end, the great African states began to weaken. This was symbolized by the death of a generation of powerful leaders: Manukosi in 1858, Sekwati of the Pedi in 1861, Mswati and Mzilikazi in 1868, Moshoeshoe in 1870, and Mpande in 1872. Just as in Germany, Italy, Canada, and Australia, so in southern Africa capitalism required the unification of communications, currencies, financial institutions, and governments. Besides, the Zulu, Swazi, and Pedi were needed as labourers, and their land—the last large fertile areas controlled by Africans—was coveted.

Governor George Grey had already proposed a federated South Africa in 1858, and in the late 1860s the discovery of gold and diamonds brought matters forward. The annexation of Basutoland in 1868 was followed by the British seizure of the diamond fields from the competing Griqua, Tlhaping, and Boers in 1871 (the Keate Award); Colonial Secretary Lord Carnarvon's more determined federation plan of 1875; Shepstone's invasion of the Transvaal in 1877; and the British invasions of Zululand and Pediland in 1879. A bid to seize Delagoa Bay was overturned by the arbitrating French president, Patrice, Count de MacMahon, in 1875; and Swaziland was left to collapse internally. With the collapse of Zulu resistance in the 1880s, the invasions of the Gaza and Ndebele kingdoms in 1893–96, and the crushing of Venda resistance in 1898, there were by 1900 no autonomous African societies left in the subcontinent. (J.R.D.C.)

Diamonds, gold, and imperialist intervention (1870–1902). Between 1870, when the diamond rush to Kimberley began, and 1902, when the South African War ended, South Africa was transformed. Midway between these dates, the world's largest goldfields were discovered in 1886 on the Witwatersrand. An economic backwater became a major supplier of precious minerals to the world economy. A scatter of disparate statelets—British colonies, Boer republics, and African kingdoms—came under British control. Predominantly agrarian societies began to urbanize and industrialize. Political responses included a nascent Afrikaner nationalism and the first modern political organization in black societies. These dramatic changes were propelled by two linked forces:

the development of a capitalist mining industry and a sequence of imperialist interventions by Britain.

Diamonds and confederation. A chance find in 1867 drew several thousand fortune seekers to alluvial diamond diggings along the Orange, Vaal, and Harts rivers. In 1870 richer finds were made in "dry diggings," and a large-scale rush followed. By the end of 1871, nearly 50,000 people were living in a sprawling, polyglot mining town that in 1873 was named Kimberley.

Initially, individual diggers, black and white, worked small claims by hand. Production was rapidly centralized and mechanized, and ownership and labour patterns more starkly divided along racial lines. Joint-stock companies bought out diggers; a new class of mining capitalists oversaw a transition from diamond digging to mining industry. By 1889 concentration became monopoly when De Beers Consolidated Mines (controlled by Cecil Rhodes) became the sole producer. While some white diggers were kept on as overseers or skilled workers, the workforce consisted mainly of African migrant workers, housed in closed compounds by the companies from the mid-1880s.

Diamonds were discovered in a zone whose sovereignty was already disputed. The Orange Free State, the South African Republic, the western Griqua under Nicolaas Waterboer, and southern Tswana chiefs all pressed competing claims. At a special hearing in October 1871, Robert W. Keate (lieutenant governor of Natal) found in favour of Waterboer. Waterboer was persuaded to request British protection against his Boer rivals, and the area was annexed as Griqualand West.

Annexation of Griqualand West

The annexation of the diamond fields signaled a more forward policy under a Liberal ministry but fell short of the ambitious confederation policy pursued by Lord Carnarvon, colonial secretary in Benjamin Disraeli's 1874 Tory government. He sought to unite republics and colonies as a self-governing federation in the British Empire. Carnarvon was influenced by Theophilus Shepstone, secretary for native affairs in Natal, who urged a coherent regional policy with regard to African labour and administration.

Carnarvon concentrated at first on persuading the Cape and the Free State to accept federation. A conference in London in August 1876 revealed how chilly these parties were to the proposal. His southern gambit frustrated, Carnarvon embarked on a northern strategy. The South African Republic (Transvaal) was virtually bankrupt, and support for President Thomas F. Burgers was dwindling. During the London conference, news arrived of a military humiliation of Burgers' forces at the hands of Sekhukhune's Pedi. Carnarvon commissioned Shepstone to annex the Transvaal. Shepstone entered the republic in January 1877 and against only token resistance proclaimed it a British colony in April.

Deft as the annexation was, administration of the new possession was maladroit. Empty coffers and insensitivity to Afrikaner resentments led to a clash over tax payments, and, under a triumvirate of Paul Kruger, Piet Joubert, and Marthinus Wessel Pretorius, the Transvaal Boers opted to fight for independence. British defeats, especially at Majuba, hastened a decision to which William Gladstone's cabinet was already inclined. Republican self-government was restored, subject to an imprecise British "suzerainty" over external relations. Confederation received its quietus with these events.

Wars of conquest. Seizure and retrocession of the Transvaal overlapped with a sequence of wars that completed the conquest of African societies. Imperial troops tipped the balance decisively against societies that had previously withstood subordination to settler control. A century of military conflict on the Cape frontier ended with the Cape-Xhosa war of 1877–78, and between 1878 and 1881 the Cape Colony defeated rebellions in Griqualand West, the Transkei, and Basutoland. Sir Bartle Frere, governor of the Cape and high commissioner for southern Africa from March 1877, rapidly decided that independent African kingdoms must be tamed if political and economic integration of the region were to become reality.

The Zulu War

Frere identified Cetshwayo's Zulu kingdom as a major obstacle to confederation. An impasse was engineered, and British and colonial troops invaded Zululand in January 1879. The annihilation of a large British force at Isandhlwana slowed the invasion, but imperial firepower ultimately prevailed. For the Zulu, political dismemberment followed upon military defeat. Divide-and-rule policies precipitated civil war in 1883, and in 1887 Zululand was annexed. British troops also took part in 1879 in a campaign that crushed Pedi military power in the northern Transvaal.

Afrikaner and African politics in the Cape. By the mid-1870s, 240,000 whites in the Cape constituted about one-third of the colony's population. Cape revenues accounted for three-quarters of the total income in the region's four settler states in 1870, and the diamond discoveries stimulated railways, public works, banking, and commerce. Although by 1870 some two-thirds of the settler population spoke Dutch or Afrikaans, political power was exercised largely by an English-speaking elite of merchants, lawyers, and landholders. In the last quarter of the century, heightened political and cultural awareness among Cape Afrikaners took organized form, and the period also saw new forms of political expression and mobilization among black voters.

The Afrikaner Bond, founded in 1880, initially represented poorer farmers and espoused an anti-British Pan-Afrikanerism in the Cape and beyond. By 1883, however, under Jan Hendrik Hofmeyr, the organization was realigned. Supported mainly by wealthier farmers and urban professionals, the Bond championed the Cape's commercial interests within a framework of regional British imperial dominance. In 1890 Hofmeyr threw his support behind Cecil Rhodes, enabling the latter to become prime minister of the Cape. The Rhodes-Hofmeyr alliance was based on their mutual desire for economic expansion northward.

A major cleavage opened between Bond politicians and English speakers loosely defined as Cape liberals. The latter grouping had material as well as ideological interests in the prosperity of an African peasantry, and in several eastern Cape electoral districts black voters were crucial in electing liberal "friends of the native." The Bond was hostile toward a commercializing black peasantry and pursued more restrictive franchise qualifications.

The piecemeal annexation of the Transkeian territories to the Cape between 1872 and 1894 greatly increased the number of Africans in the colony. Peasant production for local markets and the emergence of literate clerks and teachers enabled individuals to qualify for the vote. The rise of the Afrikaner Bond and new laws affecting franchise qualifications and taxes stimulated more vigorous African participation in electoral politics after 1884. In the eastern Cape, new political and educational bodies were created, as were the first African newspapers and African-controlled churches. The period also witnessed the first political organizations among Coloureds in the Cape and Indians in Natal and the Transvaal.

Gold mining. In 1886 prospectors established that a 40-mile belt of gold-bearing reefs existed, centring on modern Johannesburg. The rapid growth of a gold-mining industry intensified processes started by the diamond boom: immigration, urbanization, capital investment, infrastructural development, proletarianization, and labour migrancy. By 1899 the gold industry employed 109,000 people (of whom 97,000 were African migrant workers); it produced 27 percent of the world's gold and had attracted investment worth £75 million.

The world's richest goldfield was also the most difficult to work. Although abundant, the layers of gold-bearing rock ran extremely deep, and the gold content of the ore was low. To be profitable, gold mining had to be intensive and deep-level with large inputs of capital and technology. These factors ensured that production was in corporate hands almost from the outset, and amalgamation of companies proceeded rapidly. By 1898, 124 companies were arranged in nine holding companies, or "groups."

Organization of the gold industry

The group system facilitated collusion between companies to reduce competition over labour, the costs of which were crucial to their profitability. The gold mines rapidly established a pattern of labour recruitment, remuneration, and accommodation that left its stamp on subsequent

social and economic relations in the country. White immigrant miners, because of their skills, scarcity, and political power, were able to win relatively high wages. African migrants from throughout southern Africa, especially from Mozambique, were unskilled and low-paid (earning at century's end about one-ninth the wage of white miners). The mine magnates sought assiduously to limit the ratio of white to black workers and to peg migrant wages as low as possible. Migrant miners were housed in compounds, which facilitated their control and reduced overhead costs.

The road to war. Even before gold was discovered, the South African interior was an arena of tension and competition. Germany annexed South West Africa in 1884. The Transvaal claimed territory to its west, which Britain countered by creating the Bechuanaland protectorate and annexing the crown colony of British Bechuanaland. Rhodes secured concessionary rights to land across the Limpopo River, founded the British South Africa Company, and in 1890 dispatched a pioneer column to occupy what became known as Rhodesia.

While these forces jostled for position in the region at large, the domestic politics of the South African Republic became unsettled. Few 19th-century states could have adjusted with ease to the changes engendered by the gold discoveries, and certainly not a preindustrial society ruled by agrarian notables, whose president believed the world was flat. Although Paul Kruger's government made strenuous efforts to accommodate the mining industry, it was soon at loggerheads with Britain, the mine magnates, and the Uitlander ("Outlander") immigrants.

British policymakers were anxious about the Transvaal's potential as an independent actor; deep-level-mine owners chafed at corruption and inefficiency; and Uitlanders were largely excluded from the vote. In the event, Uitlander grievances provided both cause and cover for a conspiracy between British officials and mining capitalists. An Uitlander uprising in Johannesburg would be supported by an armed invasion from Bechuanaland, headed by Leander Starr Jameson, Rhodes's lieutenant, and the high commissioner would intervene to "restore order."

The Jameson Raid

The plot was botched. The Uitlander rising was called off; Jameson went ahead with his incursion in December 1895, but within days he and his force were rounded up. Rhodes had to resign as prime minister of the Cape; British Colonial Secretary Joseph Chamberlain managed to conceal his complicity. The raid polarized Anglo-Boer sentiment in South Africa, simultaneously exacerbating republican suspicions, Uitlander agitation, and imperial anxieties. The last of these had the greatest purchase, especially with the arrival in April 1897 of Sir Alfred Milner as high commissioner and governor of the Cape.

In February 1898 Kruger was elected to a fourth term as president. He entered a series of negotiations with Milner over the issue of the Uitlander franchise. Milner declared in private early in 1898 that "war has got to come" and adopted intransigent positions. The Cape government, headed by William P. Schreiner, attempted unavailingly to mediate. Marthinus Steyn, Free State president, also tried to avert war, even while he attached his cause to Kruger's. In September 1899 the two Boer republics served a last-ditch ultimatum on Britain. On its expiration on October 11, Boer forces invaded Natal.

Britain went to war to secure its hegemony in southern Africa, the Boer republics to preserve their independence. These motives were sharpened by the "new imperialism" of the late 19th century and by competing local interests. Ultimately they led to war because the discovery of gold on the Transvaal Highveld had dramatically shifted the economic centre of gravity in the region, raising the stakes and firming the resolve of the main players.

The South African War, 1899–1902. An expensive and brutal colonial war lasted two and a half years. It pitted almost 500,000 imperial troops against 87,000 republican burghers, Cape "rebels," and foreign volunteers. The numerical weakness of the Boers was offset by their familiarity with the terrain and support from the Afrikaner populace, plus the poor generalship and dated tactics of the British command. Although it was often styled a "white man's war," both sides used blacks extensively as labour, and at least 10,000 Africans fought for the British.

In the first phase of the war, Boer armies took the offensive and in December 1899 ("Black Week") punished British forces at Colenso, Stormberg, and Magersfontein. During 1900 Britain rushed reinforcements to the front; relieved sieges at Ladysmith, Kimberley, and Mafeking; and took Bloemfontein, Johannesburg, and Pretoria. In a third phase, Boer commandos eschewed conventional engagements and waged guerrilla warfare. The British commander, Lord Kitchener, devised a scorched-earth policy against the commandos and the rural population supporting them. Farms were destroyed, the countryside blockaded, and the civilian population rounded up into concentration camps. Some 25,000 Afrikaner women and children died of disease and malnutrition in these camps; 14,000 Africans died in separate camps: in Britain the Liberal leader charged the government with winning the war by "methods of barbarism."

Civilian casualties

In May 1902, Republican forces—reduced to about 20,000 exhausted and demoralized troops—sued for peace. The Treaty of Vereeniging reflected the conclusive military victory of imperial power but made a crucial concession. It promised that the "question of granting the franchise to natives [Africans]" would be addressed only after self-government had been restored to the erstwhile Boer republics. The treaty thus consigned the political fate of the black majority to the decision of white minorities.

Reconstruction, union, and segregation (1902–29). The Union of South Africa was born on May 31, 1910, parented by constitutional convention and an act of the British Parliament. The infant state owed its conception to centralizing and modernizing forces generated by mineral discoveries. Its character was shaped by eight years of "reconstruction." Between 1902 and 1910, efficient administrative structures were established, the economic dominance of gold was consolidated, and a modus vivendi was struck between Afrikaner politicians and mining capitalists. Reconstruction also ensured that settler minorities would prevail over the black majority. Prewar property relations were restored; African societies were policed and taxed more effectively, and the new constitution excluded Africans from political power. Policies proposed during reconstruction pointed toward racial segregation, which became the governing orthodoxy after 1910.

These years witnessed high levels of social and political conflict. Syndicalist white workers and Afrikaner republican diehards fought against employers and government, their clashes culminating in the Rand Revolt of 1922. New political vehicles for Afrikaner and African nationalism were constructed. Black protests against the new order ran from genteel lobbying and passive resistance to armed rural revolt, strikes, and mass mobilization.

Milner and reconstruction. In 1902 Milner transferred his headquarters from Cape Town to Pretoria. The move symbolized the centrality of the Transvaal to his mission of constructing a new order in South Africa. When Milner departed in 1905, his vision of a country politically dominated by English-speaking whites had foundered. Schemes to flood the rural Transvaal with British settlers yielded only a trickle. Compulsory Anglicization in education fanned Afrikaner nationalism instead of snuffing it out. Opposition to "Milnerism" defined the emergent political groups led by ex-Boer generals Louis Botha, Jan Smuts, and J.B.M. (Barry) Hertzog. Milner had hoped to withhold self-rule from whites in South Africa until "there are three men of British race to two of Dutch." But, when Henry Campbell-Bannerman's Liberal ministry granted responsible government to the ex-republics, Afrikaner parties—Het Volk (The People) and Oranje Unie (Orange Union) in Transvaal and the Orange Free State, respectively—won elections in 1907.

"Milnerism"

Yet, if Milner's political design failed to take shape, his blueprints for economic and social engineering were largely realized. Served by a "kindergarten" of handpicked young administrators, he made economic recovery a priority. It was imperative to restore the mines to profitability. Rail rates and tariffs on imports were lowered, and the expensive concessions granted by the Kruger regime were abolished. Milner also made strenuous efforts to ensure

cheap labour to the mines. He authorized the importation of some 60,000 Chinese indentured labourers when African migrants resisted wage cuts. Although this experiment provoked political outcries in the Transvaal and in Britain, it succeeded in undercutting the bargaining power of African workers. The value of gold production swelled from £16 million in 1904 to £27 million by 1907. Chinese miners were restricted to certain tasks, setting the precedent for a statutory colour bar in the gold mines.

The administration strove to remodel the Transvaal as a stable base for agricultural, industrial, and finance capital. Some £16 million was spent on returning Afrikaners to their farms and equipping them. Scientific farming methods were promoted, a land bank was established, and more efficient tax collection increased pressures on black peasants to work for white farmers. Central and local government bureaucracies were overhauled. Especially on the Witwatersrand, the "kindergarten" tackled town planning, public transport, housing, and sanitation. In each of these spheres, a new urban geography proceeded from the principle of separating white and black workers.

The Native Affairs Commission

The South African Native Affairs Commission (SANAC) was appointed to provide comprehensive answers to what was called "the native question." Its 1905 report proposed territorial separation of black and white landownership, systematic urban segregation by the creation of African "locations," the removal of black "squatters" from white farms and their replacement by wage labourers, and the segregation of blacks from whites in the political sphere. These (and other SANAC recommendations) provided the basis for laws passed between 1910 and 1936.

Convention and union. The new Liberal government was no more careless of British interests in South Africa than its Conservative predecessor, but it sought to secure those interests through collaboration and consent. Granting responsible government to the ex-republics was the first step in this direction. Concern in London over the electoral victory by Het Volk was short-lived. It was soon clear that Botha and Smuts accepted as axiomatic the economic preeminence of mining capital. The political corollary to this was accommodation between local and imperial political interests, expressed as a policy of "reconciliation" between Afrikaans- and English-speaking whites. The way was clear for the second stride: political unification of the region.

A constitution was drafted by a national convention, which met in Durban in 1908–09. Afrikaner leaders, and Cape Premier John X. Merriman, opted for a unitary state with parliamentary sovereignty. Four provinces enjoyed limited powers. Executive authority was vested in a governor-general, advised by a cabinet from the governing party. A technical issue, pregnant with consequences, was the delimitation of electoral constituencies so as to favour overrepresentation of rural voters. Afrikaner nationalism was the major beneficiary in subsequent years.

Two "entrenched" clauses, on language and franchise, could be amended only by a two-thirds majority vote in Parliament. Dutch and English were made official languages, reflecting the parity at the convention of Afrikaner and English-speaking delegates. The franchise issue underlined that all delegates were white males. Female suffrage was never countenanced, and imperial collaboration with settler communities was achieved at the expense of blacks. While Cape delegates favoured a colour-blind franchise, those from the Transvaal and Orange Free State demanded an exclusively white electorate. A compromise simply confirmed existing electoral arrangements. The ex-republics retained white male adult suffrage: in Natal the franchise was effectively all-white; in the Cape the vote was open to men who met property and literacy qualifications. In 1910, 85 percent of Cape voters were white, 10 percent Coloured, and only 5 percent African. Representation was further limited on racial lines: even in the Cape, only whites might stand for Parliament.

Black political responses. The South African War was fought when many black communities were hard-pressed. During the 1890s, drought and cattle disease impoverished pastoralists, and competition increased for African land and labour. During the war most black South Africans identified with the British cause, encouraged by the assertions of imperial politicians that "equal laws, equal liberty" for all races would prevail after a Boer defeat.

The Treaty of Vereeniging brusquely reneged on such promises. A sense of betrayal stimulated political protest, especially among mission-educated Africans. A multitude of organizations sprang up; their leaders sought to meet the impending union of white-ruled provinces by uniting Africans over regional and ethnic divides. While the constitutional convention deliberated, the South African Native Convention met in Bloemfontein. The first instance of coordinated action among the fledgling political associations, it was an important step toward the formation of a permanent national organization. This took place on Jan. 8, 1912, with the founding of the South African Native National Congress (later African National Congress).

Founding of the ANC

Parallel developments took place among politically conscious Coloureds. The first nationally based organization was the African Political (later People's) Organization, founded in Cape Town in 1902. Under the presidency of Dr. Abdullah Abdurahman, this body lobbied for Coloured rights and linked at times with African political groups. Indians in the Transvaal, led by Mohandas K. Gandhi, also resisted discriminatory legislation. Gandhi spent the years 1893 to 1914 in South Africa as a legal agent for Indian merchants in Natal and the Transvaal. Between 1906 and 1908, in protest against a Transvaal registration law requiring Indians to carry passes, Gandhi first implemented the methods of satyagraha (nonviolent noncompliance), which he later used with such effect in India. Not all black protest was conducted through the new middle-class organizations. In 1906 peasants in Natal refused to pay a poll tax, and their resistance developed into an armed rising, led by a chief named Bambatha. At the end of this "reluctant rebellion," between 3,000 and 4,000 Africans had been killed and 7,000 imprisoned.

Union and disunity. Louis Botha formed the first Union government on May 31, 1910, supported by the majority party in each province and by the British government. These auspices were of limited benefit as his administration entered a period of flux and violent conflict. Tensions stemmed from issues left unresolved by the constitution, from rapid but uneven economic growth and its attendant social antagonisms, and from the legacy of conquest and dispossession of indigenes by colonists.

A major axis of conflict ran between employers and organized white workers. On the Witwatersrand the Chamber of Mines and miners' trade unions were locked in combat for a decade and a half. Violent confrontations took place in 1907, 1913, and 1914. On each occasion the government deployed troops to end the strikes. White workers suspended strike action during World War I, but militancy flared again in 1919, fueled by inflation. The Chamber of Mines, squeezed by rising costs and a falling gold price, announced in December 1921 that it intended replacing semiskilled white workers with lower-paid Africans. The miners' protest stoppage in January 1922 became a general strike, and in March it developed into an armed rising, with strikers organized as commandos. Smuts, prime minister since Botha's death in 1919, used artillery and aircraft to crush the Rand Revolt, at a cost of some 230 lives. The phase of intense conflict between white unions and employers ended with the passage of the Industrial Conciliation Act in 1924, which set up new state structures for regulating industrial conflicts.

The Rand Revolt

Black workers also engaged in sporadic strikes before, during, and after World War I, and the first African trade unions emerged. In February 1920, 71,000 African gold miners struck for higher wages and lower prices, halting production for a week. Soldiers and police broke the strike, at a cost of 11 lives and more than 100 miners injured. This strike was part of a wave of protest in several cities, as inflation eroded the real wages of black workers.

In Port Elizabeth police fired on a crowd of demonstrators, killing 23 and wounding 126: Abdurahman called it "South Africa's Amritsar." Higher casualties occurred in May 1921, when troops killed 183 members of a religious sect, the "Israelites," who had occupied land at Bulhoek in the Cape and refused to leave it. In 1922 in South

West Africa (administered by South Africa as a League of Nations mandate territory from 1920), the Smuts government launched a military expedition against the Bondelzwarts people, who resisted a dog tax. More than 100 of the Bondelzwarts died. Hertzog listed these incidents in Parliament when he said of Smuts that his "footsteps dripped with blood."

Afrikaner rebellion and nationalism. Hertzog also cited the death toll of the Afrikaner Rebellion of 1914. When Britain declared war on Germany, South Africa's dominion status meant that it was automatically at war, and its troops were mobilized to invade German South West Africa. This sparked a rebellion led by former Boer generals, who were high-ranking officers in the Union Defence Force. More than 11,000 men, mainly poverty-stricken rural Afrikaners, joined the rising. The government used 32,000 troops to suppress it, and more than 300 men lost their lives in the fighting.

The rebellion, though, was a convulsive and atypical episode in the rise of Afrikaner nationalism as a political force. More telling and durable responses came from Afrikaner strata profoundly affected by economic change, war, and reconstruction. After 1902 thousands of landless families streaming into the cities indicated how the prewar rural social order had crumbled. One response to the threat of further disintegration was a "second language movement," spearheaded by teachers, clergymen, journalists, and lawyers deeply threatened by the cultural dominance of English speakers. It succeeded in its immediate aim when Afrikaans replaced Dutch as an official language in 1925; it also played a vital role in shaping Afrikaner self-awareness.

Hertzog's National Party

These forces—"poor whites" and militant intellectuals—provided much of the backing for the National Party founded by Hertzog in 1914. In the general election of 1915, the National Party won 30 percent of the poll, as Afrikaners deserted the South African Party led by Botha and Smuts. In 1920, on a platform of republicanism and separate school systems for Afrikaans- and English-speaking whites, Hertzog won a majority of both seats and votes. Smuts formed a government by allying with the strongly pro-empire Unionist Party. In 1923 the National Party entered into an electoral pact with the Labour Party. The combined voting strength of aggrieved white workers and anti-British nationalists swept Smuts from office. Hertzog headed a coalition government, known after the electoral agreement as the Pact government.

Segregation. In the first two decades of the Union, segregation became a distinctive feature of South African political, social, and economic life. It was promoted as actively under the South African Party (1910–24) as it was in the years when Hertzog was premier (1924–39). New statutes provided for racial separation in industrial, territorial, administrative, and residential spheres. It is less than helpful to explain this barrage of legislation as the product of reactionary attitudes inherited from the past. Rather, segregation represented efforts to regulate class and race relations during a period of rapid industrialization, a set of mechanisms for achieving both economic development and the maintenance of white supremacy. Indeed, the central institutions of 20th-century segregation—migrant labour, reserves, compounds, and urban locations—took shape around the gold-mining industry.

The 1911 Mines and Works Act and its 1926 successor reserved certain jobs in mining and the railways for white workers. The Natives Land Act of 1913 defined 8 percent of South Africa as African "reserves" and prohibited any purchase or lease of land by Africans outside the reserves. The law also restricted the terms of tenure under which Africans might live on white-owned farms. The Native (Urban Areas) Act of 1923 provided for segregating urban residential space and created "influx controls" to reduce access to cities by Africans. In 1926 Hertzog published bills proposing simultaneously to increase reserve areas and remove African voters in the Cape from the common roll. These aims were realized in legislation 10 years later.

The Pact years, 1924–29. The Pact government strengthened South African autonomy from Britain, aided local capital, and protected white workers against black competition. Despite the rhetoric of its election campaign, the government did little to harm the interests of the mining industry.

Hertzog played a leading role in the imperial conference that issued the Balfour Report (1926), establishing autonomy in foreign affairs for the dominions. Much parliamentary energy was consumed in wrangles over the symbols of nationalism—flag and anthem. Economic nationalism included protective tariffs for local industry, subsidies to facilitate agricultural exports, and a state-run iron and steel industry. White trade unions grew more bureaucratic and less militant, although their members enjoyed at best modest material gains. More direct benefits were enjoyed by unskilled and nonunionized whites who were aided through sheltered employment in the public sector and through prescribed minimum wages in the private sector. Although the overall level of white poverty remained high, these policies saw the manufacturing sector absorb white labour nearly twice as fast as black. In this, as in other respects, the real losers during the 1920s were Africans.

Effects of economic nationalism

For Africans, segregation meant restricted mobility, diminished opportunities, more stringent controls, and a general sense of exclusion. Economic conditions in the reserves were deteriorating, on white farms the terms of tenancy became more onerous, and the urban slums provided a harsh alternative for those who left the land. The middle-class leadership of the African National Congress (ANC) had pressed for the extension of the Cape franchise to other provinces. By 1926 it was evident that even this slender link to central political institutions was under threat of severance.

These conditions prepared the ground in which the first mass-based African political organization flourished. The Industrial and Commercial Workers Union (ICU) was until 1926 a Cape-based union with African and Coloured members drawn mainly from urban areas. It became a mass-based vehicle of rural protest, drawing scores of thousands of supporters from African tenants on white farms. The ICU linked innumerable local rural grievances with a generalized call for land and liberation. By 1929 the ICU was a spent force; unable to meet the expectations it raised in the countryside, it fell apart into several feuding factions.

The mushroom growth of the ICU stimulated radicalism in other organizations. Some ANC leaders, especially Josiah Gumede (president, 1927–30), moved leftward in the late 1920s. The Communist Party of South Africa, founded in 1921, was at first active almost solely within white trade unions, but from 1925 it recruited African members more energetically, and in 1928–29 it called for black majority rule and closer cooperation with the ANC.

These political challenges to white supremacy were reflected in the 1929 general election. For the first time since union, questions of "native policy" dominated white electoral politics. Afrikaner nationalists made "black peril" and "communist menace" their rallying cry. It was not to be the last such occasion. (C.J.B.)

The 1930s. The central theme in the history of South Africa since 1930 is the creation and eventual supersession of the most stringent system of racial segregation and discrimination that the world has known. The local whites, who never formed more than about one-fifth of the total population and were fewer than one-seventh by the 1990s, dominated the South African economy as well as the parliamentary institutions that they had inherited from Great Britain, at the expense of the Africans, Coloureds, and Indians. By 1976 white South Africans had constructed a rigid racial system known as apartheid ("apartness"), which caused untold misery and poverty and had become notorious throughout the world. The apartheid regime then began to fall apart as a result of internal resistance and fundamental changes in the distribution and exercise of power in the neighbouring states and the world beyond. By 1989 there was a stalemate. Some white political leaders realized that the regime was losing control of the country, while their opponents were aware that they lacked the means to overthrow it. There ensued a complex series of negotiations, culminating in the creation of a new, nonracial constitution, the election of a new parliament by

universal adult suffrage, and the inauguration of Nelson Mandela as president of South Africa.

In the early 1930s the South African cabinet was composed of members of the National Party, which held a majority of the seats in the House of Assembly and drew its main support from Afrikaner farmers and intellectuals. The major parliamentary opposition was the South African Party—the party of most of the English-speaking whites, who formed about 45 percent of the white population and included the major industrialists. In 1931 the Hertzog government achieved a major goal when the British Parliament passed the Statute of Westminster, which removed the last vestiges of British legal authority over South Africa. In 1934 the South African Parliament made that decision watertight in South African law by enacting the Status of the Union Act. Meanwhile, the Hertzog government had been losing support through its mismanagement of the problems created by the Great Depression, and in 1933 the prime minister decided to form a coalition with his rival Smuts, the leader of the South African Party. A year later the two organizations merged to form the United Party, with Hertzog as prime minister and Smuts his deputy.

The Statute of Westminster

The two parties and the two leaders had a common interest in favouring the enfranchised population, nearly all of whom were white, at the expense of the unenfranchised, all of whom were black. They agreed in providing massive support for white farmers; in assisting whites to rise above poverty, by providing them with jobs protected from black competition; in endorsing the mining industries' continued use of migrant African workers; in excluding Africans from participation in the conciliation machinery for settling industrial disputes; and in trying to curb the movement of Africans from the reserves into the towns. Furthermore, in 1936, with an overwhelming majority, Parliament removed from the ordinary voters' rolls those relatively few Africans who were qualified to vote; in return, it entitled Africans in the Cape Province to elect three white representatives to the House of Assembly and Africans throughout the Union to elect indirectly four white senators. It also created a Natives Representative Council with advisory powers.

After 1933, with the improvement in the international economy as the Western powers recovered from the depression, the white farmers prospered, new secondary industries were established, and South Africans of all races began to flock to the towns. South Africa was transformed from an overwhelmingly rural country, producing primary commodities for export and importing manufactured consumption goods, into a country with a diverse and nearly self-sufficient economy. However, although the standard of living of most whites improved greatly from this expansion, there was scarcely any improvement in the lives of Africans, Coloureds, and Indians. The government did add some land to the reserves, but these never exceeded 13 percent of the area of the country, and their condition was deteriorating through overpopulation and soil erosion, making it necessary for a high proportion of the men to work for wages outside the reserves, on the white farms or in the towns. There they were in an unfriendly world. African and Coloured farm labourers, scattered in small groups throughout the agricultural areas, were the most deprived of all South Africans. In the towns, life was insecure for Africans, and wages were low. In the gold-mining industry, the real wages of Africans declined by 15 percent between 1911 and 1941, when white miners were paid 12 times as much as Africans.

Role of mission schools

The education of Africans was left to Christian missions, whose resources, augmented by small government grants, enabled them to find places for only a small proportion of the African population. Missionaries did, however, run numerous schools, including some excellent high schools that took a few pupils through to the university matriculation level. Missionaries also were the dominant influence in the South African Native College at Fort Hare, which they had founded in 1916 and which included degree courses. These institutions educated a small but increasing number of Africans, who secured jobs as teachers, in the lower reaches of the civil service, or as clergy (especially in the independent churches, which had broken away from mainstream white churches). Frustrated by the fact that whites did not treat them as equals, some of them took part in opposition politics in the ANC. However, the ANC and two parallel movements—the African Political Organization (a Coloured group) and the South African Indian Congress—had little popular support and exerted scarcely any influence over the main course of events. Their leaders were mission-educated men who had liberal goals and used strictly constitutional methods, such as petitions to the authorities. The radical African ICU had collapsed by 1930, and the Communist Party of South Africa, founded in 1921, made little headway among Africans at that time.

World War II. When Britain declared war on Germany on Sept. 3, 1939, the United Party split. Hertzog proposed that South Africa should be neutral, but Smuts opted for joining Britain. Smuts's faction narrowly won the crucial parliamentary debate, the Hertzogites left the United Party, Smuts became prime minister, and South Africa declared war on Germany.

South Africa made significant contributions to the Allied war effort. Some 135,000 white South Africans fought in the East and North African and Italian campaigns, and 70,000 Africans and Coloureds served as labourers and transport drivers. South African platinum, uranium, and steel were valuable resources and, during the many months while the Mediterranean Sea was closed to the Allies, Durban and Cape Town provisioned a vast number of ships en route from Britain to Suez.

The war was an economic bonanza for South Africa. Stimulated by the reduction of imports, the manufacturing and service industries expanded rapidly, and the flow of Africans and others to the towns, already under way since 1933, became a flood. By the war's end, there were more Africans than whites in the towns. These Africans set up vast squatter camps on the outskirts of the white cities, improvising shelters from whatever materials they could find. They also began to flex their political muscles. They boycotted a Witwatersrand bus company that tried to raise fares; they formed trade unions; and, in 1946, 74,000 African gold miners went on strike for higher wages and improved living conditions.

The government suppressed that strike brutally but had no clear program for the future. White intellectuals proposed a series of reforms within the segregation framework, and the government and private industry made a few concessions, easing the industrial colour bar, increasing African wages, and relaxing the pass laws. However, the government failed to discuss these problems with black representatives. It lost credibility in the eyes of educated Africans in 1946 when it snubbed the Natives Representative Council for criticizing its handling of the miners' strike and calling for the removal of discriminatory legislation.

Afrikaner political organizations

Meanwhile, Afrikaners had created a series of ethnic organizations to promote their interests, including an economic association, a federation of Afrikaans cultural associations, and the Broederbond, a secret society of Afrikaner cultural leaders. In 1934 Daniel F. Malan, a former Dutch Reformed minister, had refused to follow Hertzog's Nationalists when they fused with the South African Party to form the United Party. Instead, he formed a new Purified National Party, which became the official parliamentary opposition. During the war many Afrikaners welcomed the early German victories—some of them committed acts of sabotage—but Malan's party adhered to constitutional methods and gradually gained support from Afrikaner clergy and intellectuals as well as from Afrikaans cultural and economic associations.

The United Party, which had won a general election in 1943 by a large majority, approached the 1948 election complacently. However, its policy statements were equivocal on race relations, while the National Party claimed that the government's weakness was threatening white supremacy and produced a statement that used the word apartheid to describe a program of tightened segregation and discrimination. With the support of a tiny fringe group, Malan's National Party won the election by a narrow margin.

Apartheid. After winning the 1948 election, the Na-

tional Party rapidly consolidated its control over the state, and in subsequent years it won a series of elections with increased majorities. In 1956 Parliament removed the Coloured voters from the common voters' rolls. To do that, the government packed the Senate with its nominated supporters to gain the two-thirds majority in a joint session of both houses required by the constitution. The 1956 law also entitled Coloured people to elect four whites to represent them in Parliament, but that arrangement did not last long. In 1969 the government abolished those seats that had represented Coloured voters, thus making the electorate exclusively white. Indians had never had parliamentary representation, and the seats of white representatives of Africans were abolished in 1959.

The Republic of South Africa

In 1961, with the approval of a majority of the white voters in a referendum the previous October (but without consultation with any Africans, Coloureds, or Indians), South Africa became a republic. The government had hoped that the country would follow a 1947 precedent, when India became a republic but continued to be a member of the Commonwealth, but, meeting criticism from other Commonwealth members, it withdrew South Africa from that loose association.

At home, the government vigorously furthered its ethnic goals. It made it compulsory for white children to attend schools that were conducted in their home language, either Afrikaans or English (except for the few who went to private schools). It advanced Afrikaners to top positions in the civil service, the army, the police, and the state corporations, such as the South African Broadcasting Corporation, which had a monopoly of radio services. It also awarded official contracts to Afrikaner banks and insurance companies. These methods raised the living standards of Afrikaners closer to those of English-speaking white South Africans.

Except for a recession during the early 1960s, the economy grew rapidly until the late 1970s. By that time, with a mixture of public and private enterprise, South Africa possessed a modern infrastructure, which was by far the most advanced in Africa: efficient financial institutions, a national network of roads as well as railways, modernized port facilities in Cape Town and Durban, and, besides the long-established diamond-, gold-, and coal-mining industries, a wide range of factories. The private sector was dominated by two great interlocking giants: the Anglo American Corporation of South Africa, founded by Ernest Oppenheimer in 1917, and De Beers Consolidated Mines. They formed the core of one of the world's most powerful networks of mining, industrial, and financial companies, employing 800,000 workers on six continents. State corporations controlled industries that were vital to national security, notably Armscor (Armaments Corporation of South Africa), which produced high-quality military equipment, and SASOL (South African Coal, Oil, and Gas Corporation), which alleviated South Africa's lack of petroleum resources by converting coal to gasoline and diesel fuel.

Program of Hendrik Verwoerd

This burgeoning economy was buoyant enough to sustain the cost of a drastic program of social engineering. The man who played the major part in transforming apartheid from an election slogan into practice was Hendrik F. Verwoerd. Born in The Netherlands, Verwoerd immigrated with his parents to South Africa when he was a child. He became a nominated senator in 1948, minister of native affairs in 1950, and prime minister from 1958 to 1966, when a deranged man assassinated him in Parliament. According to Verwoerd, the South African population comprised four distinct racial groups (white, African, Coloured, and Asian), each with an inherent culture; whites were the "civilized" group and, as such, entitled to control the state.

Parliament passed a plethora of laws to give effect to these ideas and to institutionalize the apartheid system. The Population Registration Act (1950) classified every South African by race. There were laws to prohibit interracial marriage or sex. Other laws and regulations segregated South Africans in every sphere of life: in buses, taxis, and hearses; in cinemas, restaurants, and hotels; in trains and railway waiting rooms. When a court declared that separate amenities should be equal, Parliament passed a special law to override it. Under the Group Areas Act (1950), the cities and towns of South Africa were divided into segregated residential and business areas, and the government removed thousands of Coloureds and Indians from areas classified for white occupation.

A vast bureaucracy, staffed largely by party loyalists, administered apartheid, aided by a mass of coercive laws. The Suppression of Communism Act of 1950 defined communism and its aims sweepingly and empowered the government to detain anyone it deemed likely to further any communist aims. Later laws gave the police the right to arrest and detain people without trial and without access to families or lawyers and left the courts with scarcely any means to intervene.

Africans were treated as "tribal" people, domiciled in the reserves under hereditary chiefs and bound to live there except when they were working for whites. In 1951 the government abolished the Natives Representative Council. Then it began to consolidate the scattered reserves into 8 (eventually 10) distinct territories, designating each of them as the "homeland" of a specific African ethnic community. It also manipulated homeland politics so that compliant chiefs controlled the administrations of most of those territories. Claiming to match the decolonization process that was taking place in tropical Africa, the government devolved powers onto those administrations and eventually encouraged them to become "independent." Between 1976 and 1981 four accepted independence: Transkei, Bophuthatswana, Venda, and Ciskei. However, like the other homelands, they were economic backwaters, dependent on subsidies from Pretoria, and not a single foreign government recognized them.

The "homelands"

Conditions in the homelands rapidly deteriorated, partly because they had to accommodate vast numbers of additional Africans. Attempting to reverse the flood of Africans into the towns, the government strengthened the pass laws, making it illegal for an African to be in a town for more than 72 hours without a job in a white home or business. By 1983, in a particularly brutal series of forced removals, it had ejected more than 3.5 million Africans from the towns and from white rural areas (including lands they had occupied for generations) and dumped them in the reserves.

The government also established direct control over the education of Africans. In the Bantu Education Act of 1953, it took African schools away from the missions. Then, to meet the expanding economy's increasing demand for semiskilled black labour, it created more African schools, especially in the lower grades, but subjected the students to stringent discipline and prescribed syllabi and textbooks that endorsed official policies.

A 1959 law prohibited the established universities from accepting black students, except with special permission on an individual basis. Instead, the government created five new ethnic university colleges—one for Coloureds, one for Indians, one for Zulus, and one for Sotho, Tswana, and Venda students, plus a medical school for Africans—and transformed the South African Native College at Fort Hare, which missionaries had founded primarily but not exclusively for Africans, into a state college solely for Xhosa students. It staffed these ethnic colleges with white supporters of the National Party and subjected the students to stringent controls.

Resistance to apartheid. Apartheid imposed appallingly heavy burdens on most South Africans. The economic gap between the wealthy few, nearly all of whom were white, and the poor masses, virtually all of whom were African, Coloured, or Indian, was larger than in any other country. The whites were well fed, well housed, and well cared for; Indians, Coloureds, and especially Africans suffered from widespread poverty, malnutrition, and disease. Consequently, despite the growth of the national economy, for most South Africans life was a struggle for day-to-day survival.

Nevertheless, during the 1950s the previously moribund ANC came to life under a vigorous president, Albert Lutuli, and three younger men—Oliver Tambo and Nelson Mandela, who ran a joint law practice in Johannes-

burg, and Walter Sisulu. In cooperation with the South African Indian Congress, which also had been revitalized, they organized a passive resistance campaign in 1952, when thousands of volunteers defied discriminatory laws. Three years later, in conjunction with Indians, Coloureds, and sympathetic whites, they convened a mass meeting (Congress of the People) that adopted the Freedom Charter, asserting that "South Africa belongs to all who live in it, black or white, and no Government can justly claim authority unless it is based on the will of the people." The government broke up the meeting and subsequently arrested 156 people and charged them with high treason. None was found guilty, but the trial dragged on until 1961, when the last of the accused were released.

The Freedom Charter

In 1959 a group of Africans led by Robert Sobukwe, a language teacher at the University of the Witwatersrand, believing that the alliances with white, Coloured, and Indian organizations had impeded the struggle for African liberation, broke away from the ANC and founded the Pan-Africanist Congress (PAC). On March 21, 1960, the PAC launched a fresh campaign. Thousands of unarmed Africans invited arrest by presenting themselves at police stations without passes; at Sharpeville, near Johannesburg, the police opened fire on such a crowd, killing at least 67 and wounding more than 180 Africans, most of whom were shot in the back as they were running away. Thousands of workers then went on strike, and in Cape Town 30,000 Africans marched in a peaceful protest to the centre of the city. The government reestablished control by force: it mobilized the army, outlawed the ANC and the PAC, and arrested more than 11,000 people under emergency regulations.

These events led to a change in the strategy of the congresses. Previously, they had confined themselves to nonviolent methods. After Sharpeville, however, the ANC and PAC leaders and some of their white sympathizers came to the conclusion that black people would never overcome apartheid by peaceful means alone. Violence, they concluded, was a necessary and legitimate means of resistance to the violence of an illegitimate regime. However, although their military units detonated several bombs in government buildings during the next few years, the ANC and the PAC did not pose a serious threat to the state, which had a virtual monopoly of modern weapons. By 1964 the government had captured many of the leaders, including Mandela and Sobukwe, and sentenced them to long terms of imprisonment on Robben Island in Table Bay, four miles from Cape Town. Hundreds of others fled the country. Oliver Tambo presided over an exiled ANC executive in Zambia.

A new phase of resistance began in 1973, when black trade unions organized a series of strikes for higher wages and improved working conditions. Moreover, Steve Biko and other African students founded a Black Consciousness movement that appealed to Africans to take pride in their own culture. That ideology was immensely attractive to young Africans. On June 16, 1976, thousands of children in Soweto, the African township outside Johannesburg, demonstrated against the government's insistence that they should be taught in Afrikaans rather than in English. Police opened fire, touching off a nationwide cycle of protest and subsequent repression. Once again the government reestablished control by force. Within a year, it had banned many more organizations and the police had killed more than 500 people, including Biko. Those events received worldwide execration. In 1973 the UN General Assembly had declared apartheid to be "a crime against humanity," and in 1977 the UN Security Council unanimously voted a mandatory embargo on the export of arms to South Africa.

The Black Consciousness movement

The unraveling of apartheid. By 1978 the illusion that apartheid would bring peace to South Africa was shattered. Most of the homelands were economic and political disasters. Their only significant export was labour, and most of their leaders were corrupt and unpopular. The national economy was in recession. Skilled whites were emigrating, and inflation was running high. Moreover, the global environment was changing. The Portuguese had handed over the government of Angola and Mozambique to Africans in 1974–75, and the writing was on the wall for Ian Smith's white regime in Rhodesia (which would come under African control as Zimbabwe in 1980). Increasingly isolated as the last bastion of white racial domination, South Africa had become the focus of global denunciation.

By that time, the National Party was passing under the control of a new class of urban Afrikaners—businessmen and intellectuals who, like their English-speaking white counterparts, believed that reforms should be introduced to appease foreign and domestic critics. The first attempt to give effect to their ideas occurred after Pieter W. Botha, who had been a National Party politician throughout his adult life, succeeded John Vorster as prime minister in 1978. Botha's administration applied a mixture of carrots and sticks. It repealed the bans on interracial sex and marriage; desegregated many hotels, restaurants, trains, and buses; removed the reservation of skilled jobs for whites; and repealed the pass laws. Provided that black trade unions registered, they were entitled to access to a new industrial court and permitted to strike. Also, a new constitution created separate parliamentary bodies for Indians and for Coloureds and vested great powers in an executive president, namely P.W. Botha.

P.W. Botha's reforms

However, the Botha reforms stopped short of making any real change in the distribution of power. The white parliamentary chamber could override the Coloured and Indian chambers on matters of national significance, and all Africans remained disenfranchised. The Group Areas Act and the Land Acts maintained residential segregation. Schools and health and welfare services for Africans, Indians, and Coloureds remained segregated and inferior, and most nonwhites, especially Africans, were still desperately poor. Moreover, Botha used the State Security Council—which was dominated by military officers—rather than the cabinet as his major policy-making body, and he embarked on a massive military buildup.

In 1979, in an effort to limit South Africa's economic domination of the region, South Africa's black neighbours formed the Southern African Development Coordinating Conference (SADCC), but it made little progress. Most of the export trade of the region continued to pass through South Africa to South African ports, and South Africa provided employment for 280,000 migrant workers from neighbouring countries. Botha also used South Africa's military strength to restrain those countries from pursuing antiapartheid policies. He kept South West Africa/Namibia under South African domination, sent military raids into every other southern African state, and assisted the Renamo rebels in Mozambique and the UNITA faction in its civil war in Angola.

During the 1980s the conservative British and American administrations of Margaret Thatcher and Ronald Reagan, respectively, faced increasingly vociferous pressures for sanctions against South Africa. In 1986 a high-level Commonwealth mission went to South Africa in an unsuccessful effort to persuade the government to suspend its military actions in the townships, release political prisoners, and stop destabilizing neighbouring countries. Later that year, American public resentment of South Africa's racial policies was strong enough for the U.S. Congress to pass a Comprehensive Anti-Apartheid Act over a presidential veto, banning new investments and loans, ending air links, and prohibiting the importation of many commodities. Other governments took similar actions.

Meanwhile, in 1983, 1,000 black and white representatives of 575 community groups, trade unions, sporting bodies, and women's and youth organizations launched the United Democratic Front. There followed a vast escalation of strikes, boycotts, and attacks on black police and urban councillors. Under strong pressure from white hawks, the Botha government resisted those pressures. In 1985 it declared a state of emergency in many parts of the country; a year later it promulgated a nationwide state of emergency and embarked on a savage campaign to eliminate all opposition. For three years police and soldiers patrolled the African townships in armed vehicles, destroying black squatter camps and detaining, abusing, and killing thousands of Africans, while the army also continued its forays into neighbouring countries. Rigid

The 1986 state of emergency

censorship laws tried to conceal those actions by banning television, radio, and newspaper coverage. (L.M.T.)

The brute force used by the government did not halt dissent. Long-standing critics such as Anglican Archbishop Desmond Tutu, the 1984 Nobel Peace Prize laureate, defied the government, and influential Afrikaner clerics and intellectuals withdrew their support. Resistance by black workers continued, and saboteurs caused an increasing number of deaths and injuries. The economy was severely affected, and in 1988 the army suffered a military setback in Angola, leading to the independence of Namibia in 1990. Given these circumstances, many whites came to realize that there was no stopping the incorporation of blacks into the South African political system.

Government officials held several discussions with Nelson Mandela, the imprisoned ANC leader, as these events unfolded, but Botha balked at the idea of allowing blacks to participate in the political system. An inner party coup against Botha in August 1989 forced him to step down as both party leader and president. The National Party parliamentary caucus chose F.W. de Klerk, the party's Transvaal provincial leader, as his successor. De Klerk exhibited more sensitivity to the dynamics of a world where the blatant racism that still existed in South Africa could no longer be tolerated and announced a program of radical change in a dramatic address to Parliament in February 1990; Mandela was soon released from prison. During the next year Parliament repealed the basic apartheid laws, lifted the state of emergency, freed many political prisoners, and allowed exiles to return to South Africa.

Postapartheid South Africa. *The Mandela presidency and transition to majority rule.* Nelson Mandela was elected president of the ANC in 1991, succeeding Oliver Tambo. Mandela and de Klerk, who both wanted to reach a peaceful solution to South Africa's problems, met with representatives of most of the political organizations in the country, with a mandate to draw up a new constitution. These negotiations took place amid pervasive and escalating violence, especially in the southern Transvaal and in Natal. As the bargaining continued, both Mandela and de Klerk made concessions, with the result that both of them ran the risk of losing the support of their respective constituencies. While whites were loath to forfeit their power and privileges, blacks had hoped to win complete control of the state. A majority of white voters endorsed the negotiating process in a referendum in 1992, but both white and black extremists tried to sabotage the process through various acts of terror.

Mandela and de Klerk finally reached a peaceful agreement on the future of South Africa at the end of 1993, for which they jointly received the 1993 Nobel Prize for Peace. In addition, leaders of 18 other parties endorsed an interim constitution, which was to take effect immediately after South Africa's first election by universal suffrage, scheduled for April 1994. A parliament to be elected at that time would oversee the drafting of a permanent constitution for the country. The temporary constitution enfranchised all citizens over 18, abolished the homelands, and divided the country into nine new provinces, with provincial governments receiving substantial powers. It also contained a long list of political and social rights and a mechanism through which blacks could regain ownership of land that had been taken away under apartheid.

The ANC won almost two-thirds of the 1994 vote, the National Party slightly more than one-fifth, and the Inkatha Freedom Party (IFP) most of the rest; all three received proportional cabinet representation. Mandela was sworn in as president of the new South Africa in May. Thabo Mbeki, a top official in the ANC, and de Klerk both became deputy presidents.

The new, multiparty "government of national unity" aimed to provide Africans with improved education, housing, electricity, running water, and sanitation. Recognizing that economic growth was essential for such purposes, the ANC adopted a moderate economic policy, dropping the socialist elements that had characterized its earlier programs. Mandela and his colleagues campaigned vigorously for foreign aid and investment, but capital investment entered the new South Africa slowly.

The government also had to grapple with a host of institutional problems associated with the transition to a postapartheid society. Blacks joined the civil service; antiapartheid guerrillas became members of the police and the army; and new municipal governments that embraced both the old white cities and their black township satellites sprang into existence. Labour disputes, criminal violence, and conflict between Zulu factions, especially in KwaZulu/Natal, continued. The IFP refused to participate in the process that resulted in the creation of the new national constitution that Parliament passed in May 1996. Parliament revised the constitution in October after it was reviewed by the Constitution Court; Mandela signed it into law in December of the same year.

The most important domestic agency created during Mandela's presidency was the Truth and Reconciliation Commission (TRC), which was established to review atrocities committed during the apartheid years. It was set up in 1995 under the leadership of Archbishop Desmond Tutu and was given the power to grant amnesty to those found to have committed "gross violations of human rights" under extenuating circumstances. By the time the TRC delivered its five-volume report in 1999, more than 7,000 applications for amnesty had been reviewed; of those, about 150 had been granted.

The TRC was the target of widespread criticism: whites saw it as selectively targeting them, and blacks viewed its actions as a charade that allowed perpetrators of heinous crimes to go free. Nonetheless, the TRC uncovered information that otherwise would have remained hidden. For example, details of the murders of numerous ANC members were exposed, as were the operations of the State Counterinsurgency Unit. The commission also investigated those opposed to apartheid. One of the most prominent was Winnie Madikizela-Mandela, the former wife of Mandela, who served briefly as a deputy minister in 1994–95. Her attempts to achieve other offices ended when the TRC report indicated that she had been involved in apartheid-era violence.

South Africa since Mandela. Mbeki replaced Mandela as president of the ANC at an ANC conference held in December 1996, yet Mandela's announcement that he would not seek a second term as president still came as a surprise. In the June 1999 elections, the ANC increased its share of votes to more than two-thirds, and Mbeki formed a coalition government with the IFP and the Democratic Party. (L.M.T./J.R.D.C./Ed.)

For later developments in the history of South Africa, see the BRITANNICA BOOK OF THE YEAR.

Swaziland

The Kingdom of Swaziland is a small, landlocked country embedded in the eastern flank of South Africa, where it adjoins Mozambique. It is 6,704 square miles (17,364 square kilometres) in area and extends about 110 miles (175 kilometres) from north to south and about 80 miles from west to east at its largest dimensions.

The name Swazi is the Anglicized name of an early king and nation builder, Mswati II, ruler from 1840 to 1868. The administrative centre is Mbabane, the former capital of the British colonial administration; the national capital is the seat of King Mswati III and his mother, the Ndlovukati, at Lozitha and Ludzidzini; the houses of Parliament and other national institutions are situated at Lobamba.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief and soils.* A long and complex geologic history has created a landscape with a surprising variety of relief, climate, and soils for such a small country. There are four well-defined physiographic regions, extending longitudinally from north to south in roughly parallel belts. From west to east they are the Highveld, the Middleveld, the Lowveld, and the Lubombo (Lebombo) escarpment. Geologically, the oldest formations are in the west, and the youngest are in the east.

Physiographic regions

The Highveld, covering about 30 percent of the country, is a complex of granites and more ancient metamorphosed quartzites, sandstones, and volcanics that has been eroded

into a rugged mountain land. The average elevation is between 3,500 and 4,500 feet (1,100 and 1,400 metres); the highest points are the summit massifs of Bulembu (6,108 feet [1,862 metres]) and Ngwenya (5,997 feet [1,828 metres]) in the extreme west. Known to the Swazi as Inkangala (a cold, treeless place), the Highveld was the last part of the country to be settled. Its deeper-weathered red to yellow acid soils have developed on the gentler gradients and in river valleys.

The Middleveld occupies about one-fourth of the country and has an average altitude of 2,000 to 2,500 feet. It is a region of rolling uplands and wide, well-watered valleys. It is mainly underlain by ancient granites and gneisses (metamorphosed granites), with dolerites and quartzites, which have weathered deeply to produce friable red and clay loams interspersed with shallower profiles of sands and sandy loams. To the Swazi it is known as Live ("The Country") or Inkabave ("The Navel") and is the heartland of the Swazi nation.

The Lowveld, or Bushveld, covering nearly 40 percent of the country, is a generally undulating lowland with isolated knolls and ridges rising abruptly above the general level of 500 to 1,000 feet. In general, the soils reflect the transition from the acidic granites and sandstones of the western Lowveld to the more basic basalts and dolerites of the eastern part—*i.e.*, from sandy loams in the west to red and black clays in the east, the latter being some of the most naturally fertile soils in the country. This region is called Lihlanze by the Swazi, meaning a warm place with trees—in its undisturbed state, the typical African savanna.

The Lubombo escarpment and plateau covers about 5 percent of the country, consisting of a narrow strip of about 600 square miles. It rises abruptly from the Lowveld to an average altitude of 2,000 feet, with higher peaks (Siteki and Mananga) of about 2,500 feet in the north. It is deeply dissected by the gorges of three of the main rivers that traverse the country from west to east, the Umbuluzi, the Usutu, and the Ingwavuma. The plateau soils vary considerably, from shallow sands to deeper loams, depending on the composition of the volcanic lavas that form the bedrock. The Swazi have no specific name for this part of the country.

Drainage. Swaziland is one of the best-watered countries in southern Africa. Major perennial rivers, which have their sources in South Africa, flow through the country to the Indian Ocean. They are the Lomati, the Komati, the Umbuluzi, and the Usutu. The Usutu has the largest catchment in the country, with three main tributaries, the Usushwana, the Ngwempisi, and the Mkhondvo. In the south the Ingwavuma rises in western Swaziland and also cuts through the Lubombo.

Climate. The climate is in general subtropical, but it is strongly influenced by the country's position on the eastern side of southern Africa, which exposes it to moist maritime tropical air coming off the Indian Ocean for much of the year. The cessation of maritime airflow in winter months because of intensified continental winds produces a high degree of climatic variability. The climate is also subject to steep temperature and precipitation gradients from west to east because of the fall in altitude of about 4,000 feet over a distance of about 50 miles.

Average maximum and minimum monthly temperatures are 72° F (22° C) and 52° F (11° C) in the Highveld and 84° F (29° C) and 59° F (15° C) in the Lowveld. The Middleveld occupies an intermediate position in these gradations.

Rainfall patterns

Swaziland falls within the summer rainfall region of the subcontinent, where about 80 percent of the precipitation falls during the summer months of October to March, usually in the form of thunderstorms and frontal rains. Average annual rainfall in the Highveld is about 55 inches (1,400 millimetres), in the Middleveld 34 inches, in the Lowveld about 22 inches, and on the Lubombo about 35 inches. However, variability in the annual totals is great, and figures have fluctuated dramatically from year to year. In the Middleveld, where the bulk of the population lives, the average has varied from a high of 63 inches to a low of 13 inches within a period of a few years. These extreme fluctuations appear to relate to wetter and drier than average quasi-cyclic fluctuations of from 8 to 11 years, which have been identified in the rainfall records.

Plant and animal life. The natural vegetation includes forest—confined mostly to the Highveld and the windward slopes of the Lubombo escarpment—savanna, and grassland. Factors such as soil composition and moisture produce a variety of vegetation subtypes. There are both wet and dry forests, various densities of savanna, and several grassland types, which range from sweet to sour based on their palatability when mature and dry. Altogether it is a rich flora, with ferns and flowering plants alone accounting for more than 2,600 species. Some have a very limited distribution and are found only in or around Swaziland.

The natural fauna has been severely depleted in recent years because of habitat destruction caused by the spread of the human population, and representative species such as antelope (impala, reedbuck, duiker, waterbuck, wildebeest, and kudu), hippopotamus, rhinoceros, elephant, giraffe, and zebra are found largely in protected reserves. However, smaller mammals—such as the baboon, monkey, jackal, and mongoose—may still be encountered, and several types of snake have a wide distribution. Crocodiles are also common in Lowveld rivers. Birdlife is abundant in each habitat and comprises both resident and migrant (breeding and nonbreeding) populations. The migrants come from central and North Africa and from farther afield (northern Europe and eastern Asia in the case of storks, swallows, and hawks). Distinctive among the more common birds are barbets, weavers, the various hornbills, the lilac-breasted roller, and the purple-crested loerie.

Settlement patterns. Traditionally, the Swazi lived in family homesteads (*imithi*) dispersed throughout the countryside. The only larger settlements were the homesteads of royalty and chiefs. This pattern has been modified since the late 19th century by the exposure of the rural Swazi to the money economy. Nucleated settlements grew up at important administrative and trading centres under British colonial rule from 1903, but the process of urbanization accelerated only after World War II, when the establishment of major agricultural, mining, and industrial operations acted as magnets for job seekers and created sizable company towns such as Mhlume, Simunye, Big Bend, and Mhlambanyatsi. The largest are the administrative capital of Mbabane and the commercial and industrial centre of Manzini. Some 30 percent of the population is urban, and this figure is expected to increase because the urban growth rate exceeds the growth rate of the population as a whole.

The rural population lives within a communal land tenure system administered by the traditional chiefs. A typical homestead includes the main hut of the headman (*umnumzane*); the huts of his mother, wife (or wives), and children; the kitchen and storerooms; and the cattle enclosure (*isibaya*) in front and facing east. Cattle are more than draft animals and a source of milk; they constitute a store of wealth for use on social and ceremonial occasions (*e.g.*, lobola, or bride-price).

Pattern of homestead life

The traditional pattern of homestead life is strictly seasonal. With the onset of the rains in spring (August or September), women plant gardens along the riverbanks; later, when the heavy rains come in summer (October to February), with help from the men, they plow or hoe to sow corn (maize) and sorghum (a millet) in larger fields. At this time all able women and children abandon their homesteads for the fields, and the men also join in the planting and weeding. The summer months are, on the whole, the hungry months, unless supplemented by remittances from working members of the family. Autumn to early winter (March to May) is the harvest; by July the last of the corn and sorghum has been dried and brought in. Activity then moves to the homesteads, where women and men thresh the grain, the best of which is stored and the remainder consumed at once. Winter is a time for relaxation, hunting, entertaining, and visiting. To some extent this traditional round has been disrupted by population pressure on land, by increased drift to the towns, by the absence of men working in the cities, and by the use of hired tractors for plowing, but the basic pattern is still recognizable.

The traditional centres of Swazi life are the royal villages of the *ngwenyama* (the king) at Ludzidzini and of the *ndlovukazi* (the queen mother) at Phondvo, both of which are in the "royal heart" of the country and not far from the old royal capital of Lobamba.

Ethnic composition

The people. The Swazi nation is an amalgamation of more than 70 clans. Their chiefs form the traditional hierarchy under the *ngwenyama* and *ndlovukazi*, who are of the largest clan, the Dlamini. The amalgamation brought together clans already living in the area that is now Swaziland, many of whom were of Sotho origin, and clans of Nguni origin who entered the country with the Dlamini in the early 19th century. Traditional administration and culture are regulated by an uncodified Swazi Law and Custom, which is recognized both constitutionally and judicially. The language is siSwati, which is akin to Zulu, though it shares official status with English, which is in fact used generally for official written communication.

The Swazis comprise about 97 percent of the population, the remainder being immigrants from Mozambique, South Africa, and the rest of the world. Included among these are a few thousand Europeans and Asians and their families engaged in business activities.

The majority of Swazis belong to Christian churches, both Roman Catholic and Protestant, whose missions were responsible before independence for much of the education and health services, particularly in the rural areas. However, many adherents also retain the traditional beliefs and practices of the rest of the population.

The economy. Overall, the economy displays a marked duality of large-scale intensive production and small-scale semi-subsistence activities. This produces a great contrast in incomes and living standards, which tends to be obscured by average per capita statistics. National economic policy is based on the free enterprise or market philosophy, with fiscal measures to redistribute resources to education, health, and community improvement projects. Government revenue is derived principally from receipts from the Southern African Customs Union, sales tax, and corporate and personal taxation. The budget is generally in balance, but foreign aid is a major contributor to the capital or development budget, providing a buffer to help meet any deficit in revenue. Nevertheless, the dual economy persists, and the formal employment sector is unable to absorb the annual increment of new workers generated by the country's high population growth rate. Many workers, mostly men, are forced to seek employment as migrant workers, predominantly in South Africa. Labour relations in the country are at an embryonic stage, with a generally fragmented trade union movement pitted against a longer-established employers' association and with the government endeavouring to act as referee and arbiter.

Agriculture and forestry. About two-thirds of the population lives in the rural areas, where a mixture of subsistence and commercial farming is practiced. The staple crop is corn, and other crops include sorghum (mainly for the brewing of traditional beer), pumpkins, beans, peas, and other vegetables. Crop yields are generally low, but the more progressive farmers produce on a par with the large-scale commercial sector. Because of the role of cattle as a traditional store of wealth, the livestock population, mostly cattle and goats, greatly exceeds the country's carrying capacity and is a major cause of vegetation loss and soil erosion.

Commercial farming

Large-scale commercial farming is held mainly in the hands of foreigners, although since independence the Swazi nation has acquired a large stake in this sector, especially in the largest agro-industry, the cultivation of sugarcane and the manufacture of sugar. Also of major commercial importance are the extensive man-made forests of pine and eucalyptus (in the Highveld), which supply timber to a wood-pulp mill and several sawmills. Unbleached wood pulp is the country's second largest export after sugar. The area under timber plantations is about 6 percent of the country's total area. Other important crops are citrus fruits and cotton (Lowveld), pineapples (Middleveld), rice, tobacco, and vegetables. Commercial livestock farming is also important, particularly in the Lowveld, and supports meat processing and dairy plants.

Sawmill at the foot of a man-made forest of pine and eucalyptus trees in the Highveld of western Swaziland.

Industry. Mining has declined in relative importance since the 1960s, asbestos and coal in particular. Iron ore, tin, and gold have been exploited sporadically in the past, but no mines are now active. Since 1984 diamonds have been growing in importance and are now the second largest mineral export after asbestos.

The processing of agricultural, forest, and livestock products forms the backbone of the industrial sector. Other manufactures include textiles and clothing, which expanded enormously in the 1980s, beverages, office equipment, furniture, and various other light industries.

Tourism, particularly from South Africa, has become a major sector of the economy. Centred on the hotel and casino complex in the central Ezulwini valley (about seven miles from Mbabane), the sector boasts smaller complexes at Piggs Peak in the north and at Nhlangano in the south. High-quality handmade textiles and tapestries and a range of stone and wooden handicrafts complement this sector.

Finance and trade. Swaziland, Botswana, Lesotho, Namibia, and South Africa constitute the Southern African Customs Union, which provides generally for the free movement of goods and services throughout the area. Swaziland has its own currency, the lilangeni, but is also a member of the southern African monetary union (with Lesotho and South Africa), which seeks to ensure that currencies are on par and funds move freely between the member countries.

Apart from one bank that is wholly owned by the government, the commercial banks are subsidiaries of international (including South African) banks. As a consequence of these associations, most international trade is with South Africa as part of its regional trading network. Exports are largely raw materials or lightly processed products, essentially from the agro-forestry sector, while imports consist of machinery and transport equipment, fuels and lubricants, and foodstuffs.

Transportation. Good all-weather roads link the main population centres and extend to neighbouring South Africa and Mozambique. The railway, originally constructed from the western to the eastern border for the export of iron ore through Maputo in Mozambique, has been extended to provide links to the South African network in both the north and the south of the country. The national airport is at Matsapha, about five miles from Manzini, from which the national airline (Royal Swazi

National Airways) operates scheduled services to African destinations.

Administration and social conditions. *Government.* Executive authority is vested in the king and is exercised through a dual system of government. There is a cabinet of ministers appointed by the king and presided over by the prime minister. The cabinet advises the king and is responsible to a bicameral parliament. The House of Assembly comprises 50 members, of whom 40 are elected by an electoral college and 10 are appointed by the king. The Senate has 20 members, of whom 10 are elected by the House of Assembly and 10 are appointed by the king. The electorate consists of all citizens over the age of 18 grouped into 40 constituencies (*tinkhundla*), which each elect two members to an electoral college. Elections are held at no more than five-year intervals. There are no legally recognized political parties aside from the Imbokodvo National Movement.

Dual system of government

This formal system of government is paralleled by a less formal system of committees (*amabandla*) of other prominent members of the nation who advise the king directly. In addition, the Swazi National Council advises him on all matters regulated by Swazi Law and Custom and connected with Swazi traditions and culture.

The civil service is structured into ministries, each headed by a minister appointed by the king. Other civil servants are appointed by the prime minister within the framework of a civil service board appointed by the king. There is also a judicial service commission responsible for the appointment of judges and magistrates.

The judicial system comprises the Court of Appeal, with four members appointed by the king, the High Court, headed by a chief justice, and subordinate or magistrate's courts. There are also Swazi courts, two Swazi courts of appeal, and a higher Swazi appeal court, which is subordinate to the High Court. The Swazi courts hear cases only in which all those involved are Swazi and if the charges fall within a restricted list of criminal and civil matters.

Local government is administered through four regional councils (Hhohho, Lubombo, Manzini, and Shiselweni) comprising the members of the electoral college and headed by regional administrators appointed by the king. Urban government operates under elected or appointed municipal councils (in Mbabane and Manzini) and town boards (in the smaller townships).

Land ownership

Land ownership is one of the most sensitive issues in national life. Traditionally, all land is vested in the king in trust for the nation and allocated as communal land by the chiefs. In the late 19th century, however, much of the territory was alienated as land concessions to foreigners—as owners according to them but as lessees according to the Swazi. One of the first tasks of the British crown when it assumed direct control of Swaziland in 1906 was to try to reconcile the rights of the Swazi with those of the concession holders. In 1907 it decided to reserve one-third of the country for Swazi use and to allow the concessionaires to retain two-thirds, but by World War II little progress had been achieved. The real impetus came at independence when all the crown lands became national land; shortly afterward Britain agreed to finance the repurchase of nearly one million acres. Other land was also purchased privately by the nation. Swazi Nation Land now constitutes about two-thirds of Swaziland. The remainder is held under individual title, but some of this is also under Swazi ownership, both nationally and individually.

Education. Schooling was introduced as a part of missionary activity in precolonial times, and missionaries continue to influence the education system. The Swazi nation itself set up schools as early as 1906, and a number of chiefs established what were known as "tribal" schools. However, it was only after independence that the coverage of primary and secondary schools began to increase dramatically and to enable more than four-fifths of the school-age population to attend full-time. As a result, illiteracy is declining steadily. State education is not free, and school fees constitute a major financial commitment for parents. There are also teacher-training and vocational and industrial training centres, as well as a university.

Health and welfare. The initial stimulus for health services came from church missions and from industrial establishments catering to large numbers of employees and their dependents. They established both hospitals and rural clinics. There are also private medical practitioners in all the larger urban centres. Chief causes of illness are intestinal infections, tuberculosis, food deficiencies, and respiratory diseases. After its virtual elimination in the 1950s, malaria has again become a major disease, especially in the Lowveld, where there has been a large influx of infected immigrant labour from Mozambique. By 2000, Swaziland suffered from one of the highest infection rates of HIV in the world, with nearly one-fourth of the population being afflicted.

Cultural life. Tradition continues to play an important role in Swazi society, both at the national ceremonial level and in day-to-day personal contacts. This reflects the unity of the Swazi as one nation under a traditional leader and especially their reverence for the struggle of King Sobhuza II over the 61 years of his reign to regain their independence.

The Incwala and Umhlanga

The two main cultural events are the Incwala in December and the Umhlanga in August. The Incwala is sometimes described as a first-fruits ceremony, but, spread over six days, it is a much more complex ritual of renewing and strengthening the kingship and the nation, with songs and dances used only on this occasion. The Umhlanga, or Reed Dance, brings together the maidens of the country to cut reeds for the annual repairs to the windbreaks of the queen mother's village; it lasts for five days. It is also symbolic of the unity of the nation and of its perpetuation through the massed ranks of young women. Both ceremonies are held at the national capital of the queen mother.

Other ceremonies are associated with the communal weeding and harvesting of the king's fields (and those of the chiefs) and with customary marriages. Most ceremonies are accompanied by traditional music, songs, and dancing. Musical instruments are simple in design, a kudu horn (*impalampala*) used for hunting or herding cattle, a calabash attached to a bow (*umakweyane*) for love songs, the reed flute, played by small boys while herding, and rattles made of seedpods attached to the wrists and ankles. However, more typical of the homestead nowadays are the radio and record and tape players.

For statistical data on the land and people of Swaziland, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Early history. The Swazi nation is a relatively recent political grouping, the main amalgamation of clans having taken place under Dlamini military hegemony about the middle of the 19th century. However, the record of human settlement in what is now Swaziland stretches far back into prehistory. The earliest stone tools, found on ancient river terraces, date back more than 250,000 years, and later stone implements are associated with evidence of *Homo sapiens,* perhaps as long ago as 100,000 years. By 42,000 years ago the inhabitants were quarrying red and black hematite ore for cosmetic purposes on the top of the Ngwenya massif (where in 1964 a large opencut mining operation was developed to exploit the rich ore deposit). This ranks as one of the world's earliest known mining and trading activities, and mining continued for many thousands of years after that. Much later—about 20,000 years ago—the archaeological record reveals occupation by the ancestors of the San hunter-gatherers, who created the distinctive rock paintings found throughout the western part of the country.

About 2,000 years ago groups of Bantu-speaking peoples (Nguni, Sotho, and Tswana) moved southward across the Limpopo River. They cultivated crops, kept livestock (sheep and goats), used pottery, and smelted iron—hence their designation as Early Iron Age peoples. At a later date cattle were introduced. These people are recorded at Ngwenya, where the mining of iron ore has been dated to about AD 400. During the following centuries the more attractive areas of Swaziland were settled by these ancestors of the Nguni and Sotho clans, whom the Swazi encountered in the late 18th and early 19th centuries.

The ancestors of the Dlamini clan were part of this southward movement, which reached the Delagoa Bay area (now Maputo) of Mozambique some considerable time before the arrival of the Portuguese in the early 16th century. There they settled as part of the Thembe-Tonga group of peoples until the mid-18th century, when, probably because of dynastic conflict, they moved southward along the coastal plain between the mountains and the Indian Ocean, "scourging the Lubombo" as a royal praise song puts it. Up to this time they called themselves Emalangeni, after an ancestral Langa. Later they moved westward through the Lubombo range and up the Pongola valley, where in about 1770 under their king Ngwane III they established the first nucleus of the Swazi nation (bakaNgwane) near what is now Nhlangano.

Emergence of the Swazi nation. This was a turbulent period in the history of southeastern Africa, when a number of major clan groupings were struggling for supremacy. Two of these, the Ndwandwe and the Zulu, located to the south of the new Ngwane homeland, constituted a serious threat to the Dlamini, who strove to establish their control over the clans among whom they had settled. Nevertheless, by the end of the century, they had achieved considerable success in assimilating some of these clans and in forging bonds with others to create a new political grouping. However, this new power base was not strong enough to ward off aggression by their southern neighbours, so about 1820 under their new king—Sobhuza I, or Somhlohlo ("The Wonder")—they moved northward to establish a safer heartland in central Swaziland (the Middleveld). There the Dlamini consolidated their power under Sobhuza I and his son Mswati II. Part of this success must be attributed to Sobhuza's adoption of the Zulu age-group system of military organization, which created regiments across clan loyalties and was strictly disciplined. By 1860 they had extended their power through conquest and assimilation far beyond present-day Swaziland under Mswati II, whom later generations described as "their greatest fighting king" and who gave his name to the nation.

Consolidation of power

At the peak of their power, however, a new factor had emerged in the regional geopolitics, which over the next 40 years caused the gradual contraction of Swazi territorial and political authority. This was the expanding Boer republic of the Transvaal and the growing British imperial presence, especially after the discovery in South Africa of diamonds in 1867 and gold in 1871.

The main destabilizing force was the stream into the country of European prospectors and concession hunters, which the Swazi were able to contain for a while but which became a flood after the kingship passed to Mbandzeni in 1875. By 1890 so many concessions had been granted for so many purposes (in addition to land and mineral rights) that practically the whole country was covered two, three, or even four deep in concessions of all kinds and for different periods. Although the Swazi maintained that these were all leasehold rights that would terminate at some future date, they had, as it later transpired, signed away their independence.

The European presence

In 1888 the Swazi tried to regulate the new influences that the influx of Europeans had created by granting them a charter of conditional self-government subject to the royal veto. Behind the concessions scramble by individuals, however, lay the intrigue and conflict of the two white powers, the Boers and the British. The former needed a route to the sea, while the latter wanted to contain them. Swaziland stood in the way, as an obstacle to be manipulated by both.

In 1890, under a convention between the British government and the South African Republic, a provisional government consisting of representatives of the two powers and a representative of the Swazi people was set up. In 1893 the British government signed a new convention permitting the South African Republic to negotiate with the Swazi regent and her council for a proclamation allowing the republic to assume powers of jurisdiction, legislation, and administration without the incorporation of Swaziland into the republic. The Swazi refused to sign the proclamation, but in 1894 another convention was signed by the two powers, virtually giving unilateral effect to its terms. After the South African War of 1899–1902 all the rights and powers of the republic passed to Great Britain, and in June 1903, by an order in council under the Foreign Jurisdiction Act, the governor of the Transvaal was empowered to administer Swaziland and to legislate by proclamation. In 1906 these powers were transferred to a high commissioner for Basutoland, Bechuanaland, and Swaziland.

Colonial administration. The colonial years from 1906 to the late 1940s saw Swaziland drift into a backwater of the British Empire. A fundamental reason was that provision had been made in the South Africa Act of 1909 (which established the Union of South Africa as a British dominion) for the possible eventual transfer of Swaziland (and Basutoland and Bechuanaland) to the union. While this possibility existed, no socioeconomic improvement took place, and it was difficult to distinguish Swaziland from the neighbouring rural areas of South Africa. Politically, the situation was epitomized in the downgrading of the title of king to that of paramount chief and of his function to that of "native administration." Despite a number of requests from South Africa, however, the imperial power declined to transfer Swaziland. This resolution was stiffened by events in South Africa after the 1948 election, which heralded the onset of apartheid. Also, from 1945 onward, Britain had begun to tackle socioeconomic problems. By the mid-1950s the issue of transfer was dead, though the grand apartheid design of separate homelands for Africans still included Swaziland.

From 1960 the economy forged ahead steadily, but sociopolitical progress followed more slowly. A constitution providing for limited self-government was promulgated in 1963, and in 1967 the country became a protected state under which the kingship was restored. This was followed by full independence on Sept. 6, 1968. (Jo.R.M.)

Independence

Swaziland since independence. King Sobhuza II had been installed as the *ngwenyama* of the Swazi nation in 1921. After 1968 the governing system moved from being a constitutional monarchy to being directly controlled by the king. Five years after independence, Sobhuza repealed the constitution designed by the British and restored the traditional system of government, in which all effective power devolved to the royal capital. A system of local government, known as the *tinkhundla*, allowed traditional authorities to control the government at the local level. The king's concession to modern government was to retain the cabinet system with a prime minister and other ministers, although they were all chosen by the king. Under his rule, Swaziland enjoyed a remarkable degree of political stability and economic progress. Emphasis was placed on education—which had been neglected in colonial times—on health, and on other human-resource developments. A major concern, however, was a growing economic dependence on South Africa.

King Sobhuza II died on Aug. 21, 1982; his death was followed by a power struggle within the royal family that was not resolved until 1986, when the teenage heir, Prince Makhosetive, was installed as King Mswati III. Under his leadership, Swaziland continued to follow a traditional form of government, but throughout the 1990s opposition continued to grow. The king responded by saying he would give the people a greater voice in government, but he failed to provide much substance by the beginning of the 21st century. Political parties were still not allowed, and a promised new constitution failed to materialize. Swaziland's biggest 21st-century problem was the rising number of cases of AIDS. (Ed.)

For later developments in the history of Swaziland, see the BRITANNICA BOOK OF THE YEAR.

Zambia

Zambia is a landlocked republic in southern Africa with an area of 290,586 square miles (752,614 square kilometres). The capital is Lusaka. Zambia's population is highly urbanized, and large parts of the country are thinly populated. Population is concentrated in the "Line of Rail," the area served by the railway linking the Copperbelt with

Lusaka and the border town of Livingstone. Zambia has a long land border on the west with Angola but is divided from its neighbours to the south by the Zambezi River. To the southwest is the thin projection of Namibian territory known as the Caprivi Strip, at the eastern end of which four countries (Zambia, Namibia, Botswana, and Zimbabwe) appear to meet at a point—a "quadripoint"—although the precise nature of the meeting is contested. Man-made Lake Kariba now forms part of the river border with Zimbabwe. Mozambique is Zambia's neighbour to the southeast, Malaŵi to the east, and Tanzania to the northeast. The long border with Congo (Kinshasa) starts at Lake Tanganyika, crosses to Lake Mweru, and follows the Luapula River to the Pedicle, a wedge of Congolese territory that cuts deep into Zambia to give the country its distinctive butterfly shape. Westward from the Pedicle the frontier follows the Zambezi-Congo watershed to the Angolan border. The country's name is derived from the Zambezi River, which drains all but a small northern part of the country.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Most of Zambia forms part of the high plateau of this part of Africa (3,000 to 5,000 feet [900 to 1,500 metres] above sea level); major relief features occur where river valleys and rifted troughs, some lake-filled, dissect its surface. Lake Tanganyika lies some 2,000 feet below the plateau, and the largest rift, that containing the Luangwa River, is a serious barrier to communications. The highest elevations occur in the east, where the Nyika Plateau on the Malaŵian border is generally over 6,000 feet, rising to more than 7,000 feet in the Mafinga Hills. The general slope of the plateau is toward the southwest, although the drainage of the Zambezi turns eastward to the Indian Ocean. Over most of the country, ancient crystalline rocks are exposed, the product of prolonged erosion processes. In western Zambia they are overlain by younger sandy deposits, relict of a once more extensive Kalahari desert. In central and eastern parts of the country, downwarping of the plateau surface forms swamp- or lake-filled depressions (*e.g.,* Lake Bangweulu, the Lukanga Swamp); in more elevated regions, ridges and isolated hills made up of more resistant rocks punctuate otherwise smooth skylines.

Highlands and rifts

Drainage. The continental divide—between the Congo River drainage, which flows to the Atlantic, and that of the Zambezi, which drains into the Indian Ocean—runs along the Zambia-Congo (Kinshasa) border west of the Pedicle and then northeastward to the border with Tanzania. Both the Luapula (which drains the Bangweulu basin into Lake Mweru) and Lake Tanganyika are tributary to the Congo. The rest of the country lies within the Zambezi basin, the river itself rising in northwestern Zambia and circling through Angola before traversing the sandy plains of western Zambia. At the Victoria Falls it drops 300 feet into a milewide chasm at the head of the gorge leading down to Lake Kariba and the troughlike middle part of its valley. It has two main tributaries in Zambia. Rising on the Copperbelt, the Kafue River drains the Lukanga Swamp and Kafue Flats before an abrupt descent to the Zambezi. The Luangwa River, mostly confined within its rift trough, is quite different. The Bangweulu Swamps and the Kafue Flats are wetlands of international importance.

Geology and soils. The oldest rocks of the country are volcanics and granites of the Bangweulu block in the northeast. These are 2.5 billion years and older and have been unaffected by orogenic processes since Precambrian times. This old structure is partly covered by ancient sedimentary rocks, and together they constitute the basement complex. Sedimentaries of the Katangan System (550 to 620 million years old) are extensive in the central areas, and mineralization of these rocks is the basis of Zambia's mining industry. Later sedimentary rocks of the Karroo (Karoo) System filled rifted troughs in the plateau surface, some of which, as in the Luangwa and middle Zambezi valleys, have been partially re-excavated. Coal seams occur in Karroo rocks to the north of Lake Kariba. These structural troughs are ancient features. Younger rifts in the north, part of the East African Rift System, are occupied by Lakes Mweru and Tanganyika. Karroo and older sedimentaries are also found in the west, buried under the predominantly sandy deposits of the Kalahari System.

The soils of the plateau are generally of poor quality, long-continued weathering and erosion having leached many of their nutrients. Much of the plateau is covered by the so-called Sandveld soils, which have a sandy surface layer overlying a clayey subsoil, often with laterite (an iron-rich horizon). Shifting cultivation is widespread, and for more permanent cultivation soils need to be carefully managed. More fertile red clay soils occur over limestone and basic rocks and have attracted commercial farming. Soils of the Kalahari Sands have little agricultural potential and are mainly under woodland. The black clay soils of some floodplains and swamp areas are highly fertile but difficult to cultivate, being waterlogged in the rainy season and rock-hard when dry.

Sandveld soils

Climate. Although Zambia lies within the tropics, its climate is modified by the elevation of the country and is generally favourable to human settlement and comfort. The marked seasonal pattern of rainfall is caused by the north and south movement of the intertropical convergence zone (ITCZ), following the apparent movement of the Sun. In January the ITCZ is in its southernmost position, and the rainy season is at its peak; by June it has moved north, and the weather is dry. Summer rains reduce the high temperatures that might be expected at this time.

Rainfall (concentrated in just five months) is highest over the Bangweulu basin (more than 60 inches [1,500 millimetres] annually) and along the Congo-Zambezi watershed, declining southward to the middle Zambezi valley, which averages less than 28 inches. The Luangwa valley is also drier than the surrounding plateau. Rainfall is less reliable in the drier regions, and failure of the rains in the south and southwest periodically brings famine to these areas.

Temperature is modified by elevation, mean daily maximum temperatures higher than 100° F (38° C) occurring only in the Luangwa valley and the southwest. The coolest area is the high Nyika plateau on the border with Malaŵi. During the cold months (June and July), the area west of the Line of Rail is coolest, with mean minimum temperatures mostly under 45° F (7° C). Sesheke, in the southwest, has frost on an average of 10 days per year.

Average annual hours of sunshine range from more than 3,000 in the southwest to less than 2,600 on the eastern border. Winds are predominantly easterly-southeasterly, although in the rainy season winds blow from the northwest and north. Wind speeds are rarely strong enough to cause damage.

Although the major contrast is between the rainy season and the drier months, three seasons may be identified.

Three seasons

The warm wet season lasts from November until April. The movement into Zambia of the moist Congo air mass from the northwest heralds the start of the rains, in the north usually in early November and toward the end of the month around Lusaka. The change from dry to wet conditions is transitional rather than abrupt. December and January are the wettest months. Cloud cover lowers maximum temperatures but also limits radiative heat loss at night, so that minimum temperatures are kept relatively high. Relative humidity values are high, typically 95 percent in early morning but declining to 60–70 percent by midafternoon. Sunshine is surprisingly frequent, Lusaka averaging six hours of sunshine per day in January. Rainfall declines rapidly in April with the northward movement of the ITCZ.

The cool dry season lasts from April until August. The sun is overhead in the Northern Hemisphere, so temperatures are low; July is usually the coolest month. Clear skies allow maximum radiation and result in especially low temperatures on calm nights, with occasional ground frost occurring in sheltered valleys.

The hot dry season lasts from August until November. This is a period of rapidly rising temperatures; just two months separate July, the coldest month, and October, usually the hottest (although if the rains are delayed November can be hotter). Usually by mid-October cooler oceanic air moves in, leading to increasing humidity and cloud formation. High temperatures and increasing hu-

midity make this one of the least comfortable times of the year, although the first rains wash away dry-season dust.

Plant and animal life. On the plateau, miombo woodland is characteristic: a semicontinuous tree cover dominated by small leguminous trees of the *Brachystegia* and *Julbernardia* genera but with a significant grassy undergrowth. Burning of the grasses in the dry season causes the trees to develop a corky, fire-resistant bark. Mopane woodland, in which *Colophospermum mopane* dominates but in which the baobab is distinctive, occurs in the drier and hotter valleys of the Zambezi in the south and in the Luangwa valley. Zambezi teak (*Baikiea plurijuga*) occurs in the southern fringe of the area covered by the Kalahari Sands. Mukwa (*Pterocarpus angolensis*), a good furniture timber, is found in the Lake Bangweulu area. More than 8 percent of the country has been set aside as forest reserve or protected forest areas.

Where there is seasonal flooding—in swamps and on floodplains, notably in the Bangweulu and Lukanga regions, in the upper Zambezi, and on the Kafue Flats—open grasslands are characteristic. On the plateau the tree cover is broken by grass-covered *dambos:* shallow saucer-shaped valleys, often lacking a surface watercourse.

The variety of Zambia's mammals is notable, and there are large concentrations on the major floodplains, particularly in the national parks of the Luangwa and Kafue valleys. Depletion of wildlife has occurred because of the spread of human activities outside the parks, while poaching is a serious threat within. The illegal trade in rhino horn has been responsible for the virtual elimination of the rhinocerous from Zambia, and poaching of elephants for their tusks has greatly reduced their numbers, despite government measures against it. There are a large range of smaller mammals and varied and numerous birdlife. The fish eagle, Zambia's national emblem, is common on large stretches of open water.

Reptiles include crocodiles, tortoises, terrapins, a variety of lizards, and many poisonous and nonpoisonous snakes. Insects of most orders are prevalent. Termite mounds, often large and sometimes pinnacled, are a landscape feature of some areas and can hinder farming operations.

Wildlife is protected in 19 official national parks and 34 game-management areas, which together constitute more than one-third of the country. Eight national parks have tourist facilities: South and North Luangwa, Kafue, Lochinvar, Blue Lagoon, Sumbu, Nyika, and Mosi-oa-Tunya. Kafue, the oldest and largest of these parks (8,650 square miles), is on the plateau and has generally low game concentrations, although it is noted for the variety of species of antelope it hosts. Lake Itezhi-Tezhi, a reservoir behind a regulating dam on the Kafue, has flooded part of the park. South Luangwa (3,500 square miles) has one of Africa's largest (but declining) elephant concentrations. North Luangwa, a park only recently made accessible to tourists, offers true wilderness adventure: walking safaris. Thornicroft's giraffe is unique to the Luangwa valley. The other parks are much smaller. Lochinvar, on the Kafue Flats, is of particular interest to bird-watchers, with more than 400 species recorded. The Kafue lechwe, a type of aquatic antelope, is unique to the flats. The Mosi-oa-Tunya National Park protects the environs of the Victoria Falls. Nyika National Park was established to preserve remnant patches of montane forest. Sumbu, on the shores of Lake Tanganyika, is renowned for easy sightings of the rare sitatunga.

National parks

The people. *Ethnic and linguistic composition.* Relative to the country's area, Zambia's population is small (although, with a growth rate of more than 3 percent per annum, increasing rapidly). It is highly urbanized, with well over half the population living in the four provinces along the Line of Rail. The movement of people from the rural areas into the towns was particularly marked after independence. Government efforts to reverse the flow have had only limited success.

Most Zambians speak Bantu languages and are descended from farming and metal-using peoples who settled in the region over the past 2,000 years. Cultural traditions in the northeast and northwest indicate influences and migrations from the upper Congo basin. There are also some descendants of hunters and gatherers who seem to have been pushed back into the Kalahari, the Bangweulu and Lukanga swamps, and the Kafue Flats. In the 19th century invaders arrived from the south: Ngoni settled in the east, while the Kololo briefly ruled the Lozi in the upper Zambezi valley. Europeans began to enter in significant numbers in the late 19th century.

Although most Zambians are of Bantu origin, the complex patterns of immigration have produced wide linguistic and cultural variety. Eighty different languages or dialects have been identified in Zambia; they can usefully be considered as comprising 14 groups, of which the Bemba group is the most widespread, accounting for more than one-third of the population. Second in importance is the Nyanja group (about 17 percent), while the Tonga group is about 15 percent.

The non-African population tends to be located in the towns, although the commercial farming community, found mainly in the central and southern regions, includes Europeans and whites from South Africa, some holding Zambian citizenship. Many Europeans left at indepen-

A Thornicroft's giraffe in the mopane woodland of the Luangwa valley (North Luangwa National Park), eastern Zambia.

dence, and their numbers have steadily declined, partly owing to regulations that restrict the employment and residence of nonnationals.

By contrast, the number of Asians has risen since independence. The majority are engaged in the retail trade, and they also are concentrated in the major towns, because in 1970 non-Zambians were prohibited from trading in rural areas. Most are Indians, mainly Gujarātī speakers from western India.

Ethnic distribution

There is some relationship between the distribution of major tribal groups and the administrative division of the country into provinces. The Western (formerly Barotse) Province is dominated by the Lozi, who live on and about the floodplain of the upper Zambezi. Lozi society is markedly centralized under the leadership of a king, the *litunga,* and at one time nurtured separatist aspirations.

In the North-Western Province, adjoining the Angolan and Zairean borders, there is no single dominant group; the peoples here include the southern Lunda and the Luvale, Chokwe, Luchazi, Mbunda, Ndembu, and Kaonde.

The Southern Province contains the Ila-Tonga peoples, of which 12 separate groups can be identified, speaking closely related dialects. Settlement is characterized by dispersed homesteads, and there are no chiefs. Traditionally cattle-owning, they occupy an area of above-average soil fertility through which the railway was built, encouraging early involvement in commercial agriculture. Migration to urban areas is of lesser importance there than elsewhere.

The Northern Province is dominated by the Bemba, who formed an extensive kingdom in the 19th century. The province was a major source of mine labour, and Bemba has become the lingua franca of the Copperbelt as well as the most widely spoken language in the country. Most languages in the northeast of the province are closely related to languages in Tanzania and Malaŵi.

Luapula Province extends along the river of that name from Lake Bangweulu to Lake Mweru and is inhabited by a number of Bemba-speaking but culturally distinct peoples (among them the Lunda, Kabende, Aushi, and Chishinga). Fishing is the major economic activity. In the 19th century the valley was dominated by the Lunda kingdom of Kazembe.

The Eastern Province is the home of the Nsenga, Chewa, Kunda, and Ngoni. The last invaded from the south during the 19th century but took the language of the peoples that they raided. The dominant language is Nyanja, which is also spoken in Malaŵi and is the lingua franca in Lusaka, to which many migrants from this area have moved.

The ethnic boundary between the Ila-Tonga and the Lala-Lamba groups runs approximately through the Central Province, with the Lenje-Soli peoples occupying a buffer area between the two. The Lenje are related to the Ila-Tonga, and the Soli to the Lala-Lamba, who, in turn, are connected with the Kaonde of the North-Western Province.

The Copperbelt (formerly the Western) Province is the location of the mining industry. The population is composed of people from all parts of Zambia, as well as some from neighbouring countries. This is true also of Lusaka Province, a small province created around the capital from the southern part of Central Province in 1976.

There are seven official vernacular languages: Bemba, Nyanja, Lozi, Tonga, Luvale, Lunda, and Kaonde, the latter three being languages of the North-Western Province. English is the official language of government.

Dominance of Christianity

Religion. Zambia is predominantly a Christian country, although few have totally abandoned all aspects of traditional belief systems. The first Christian missions arrived before colonial rule, and the growth of adherents was greatly assisted by the schools that they established. The Roman Catholic church is today the largest single denomination, but Anglicans, Baptists, Methodists, and others are well established. The growth of fundamentalist churches has been particularly noticeable since independence, and the government of the newly independent country soon ran into conflict with two of these, the Jehovah's Witnesses and the Lumpa church. The Asian community is predominantly Hindu, the rest mainly Muslim. There are relatively few Muslims in the African population.

The economy. Zambia's economy is heavily dependent on mining, in particular the mining of copper. Unfortunately, reserves of copper ore at some mines are becoming depleted, costs of production have increased, and the price of copper on the world market has slumped. There is thus a great need to broaden the base of the economy. Agriculture is relatively poorly developed, however, and major investment in manufacturing industry did not take place until after independence. State involvement in all aspects of the economy has been a feature of independent Zambia and has created a highly centralized and bureaucratic economic structure, although changes in the political structure of the country in the early 1990s were accompanied by efforts to increase private investment and involvement, particularly in the industrial sector.

Shortly after independence, Zambia embarked on a program of national development planning, the Transitional Development Plan, preceding the First National Development Plan of 1966–71. This later plan, which provided for major investment in infrastructure and manufacturing, was largely implemented and generally successful (which was not true of subsequent plans).

Post-independence restructuring

A major switch in the structure of the country's economy came with the Mulungushi Reforms of April 1968, in which the government declared its intention to acquire an equity holding (usually 51 percent or more) in a number of key foreign-owned firms, to be controlled by the Industrial Development Corporation (INDECO). By January 1970 a majority holding had been acquired in the Zambian operations of the two major foreign mining corporations, the Anglo American Corporation and the Rhodesia Selection Trust (RST), which became the Nchanga Consolidated Copper Mines (NCCM) and Roan Consolidated Mines (RCM), respectively. A new parastatal body, the Mining Development Corporation (MINDECO), was created. Government control was later extended to insurance companies and building societies, which were placed within a new parastatal body, the Finance and Development Corporation (FINDECO). The banks successfully resisted takeover. INDECO, MINDECO, and FINDECO were brought together in 1971 under an omnibus parastatal, the Zambia Industrial and Mining Corporation (ZIMCO), to create one of the largest companies in sub-Saharan Africa. In 1973 management contracts under which the day-to-day operations of the mines had been carried out by Anglo American and RST were ended. In 1982 NCCM and RCM were merged into the giant Zambia Consolidated Copper Mines Ltd. (ZCCM).

Programs of nationalization, particularly of the mining industry, were ill-timed. The massive increase in the price of oil in 1973 (which greatly inflated the import bill) was followed by a slump in copper prices in 1975 and a diminution of export earnings. The price of copper, which in 1973 accounted for 95 percent of all export earnings, halved in value on the world market in 1975. By 1976 there was a balance-of-payments crisis, and the country became massively indebted to the International Monetary Fund (IMF). There was little hope of putting the proposals of the Third National Development Plan (1978–83) into effect: crisis management, not long-term planning, was the reality.

By the mid-1980s Zambia had become one of the most indebted nations in the world relative to its gross domestic product (GDP). As the price for its continuing support, the IMF was able to insist that the Zambian government introduce programs aimed at stabilizing the economy and restructuring to reduce dependence on copper. Measures included ending price controls, currency devaluation, reductions in government expenditure, the ending of subsidies on food and fertilizer, and increased prices for farm produce. The removal of food subsidies caused massive increases in the price of basic foodstuffs and led to rioting. Unable to cope with internal opposition to the new policies, Zambia broke with the IMF in May 1987, introducing its own New Economic Recovery Programme in 1988; it subsequently moved toward a new understanding with the IMF in 1989. In a major policy turnabout in 1990, reflecting events in eastern Europe and the Soviet Union, the intention to partially privatize the parastatals

was announced. The new government of the Movement for Multiparty Democracy, which came into power in November 1991, promised to liberate the economy and introduce a free-market system.

The union movement, especially among the mine workers, has been a strong influence on political development since the 1930s, and President Frederick Chiluba, who succeeded the country's first president, Kenneth Kaunda, in 1991, had been leader of the trade union movement.

Copper and other minerals

Resources. Copper, the basis of Zambia's prosperity in the first decade of independence, is a declining asset. Alternatives such as optical glass fibre have reduced market demand, and exhaustion of reserves in existing mining areas has led to increased costs. Large-scale mining could end by 2005, although small-scale operations will continue long after that. Cobalt occurs in association with copper. Lead and zinc mining at Kabwe began in 1906, predating the large-scale mining of copper. Underground mining at Kabwe has practically ended, although reworking of mine dumps has prolonged activity at the mine.

Other minerals worked in Zambia include gold and silver, both of which occur in association with copper. Iron ore is found near Mumbwa. There is an increasing awareness of the value of Zambia's gemstones. Emeralds, mined near Luanshya and Ndola, are cut and polished locally. Amethyst, aquamarine, and tourmaline are also mined. Large deposits of cosmetic-grade talc are found near Ndola and Lusaka. Limestone is widely found on the Copperbelt and in the Lusaka district and is quarried for stone, lime, and cement; associated with it are workable occurrences of marble.

The country once relied on coal carried by rail from Hwange in what is now Zimbabwe, but, following Rhodesia's Unilateral Declaration of Independence (UDI) in 1965, relatively poor-grade coal deposits were developed at Maamba in the Gwembe area, adjacent to Lake Kariba. Although there has been extensive prospecting for oil in the Karroo sediments of the middle Zambezi, the Luangwa, and the southwest, the search has so far been unsuccessful. Nor has prospecting for uranium discovered workable quantities of the ore.

Hydropower

There is, however, one rich energy source: hydropower. Large rivers descending from the plateau into the rifted troughs of the Zambezi provide scope for hydropower development, and a major gorge on the middle Zambezi enabled it to be dammed to form Lake Kariba, the world's largest man-made lake of its time. The first power station at Kariba was built on the south side of the river, but a 600-megawatt station on the Zambian side was completed in 1977, shortly after the completion of a 900-megawatt station in the Kafue Gorge, south of Lusaka. There is an earlier power station at the Victoria Falls. Another dam on the Zambezi, which would need the collaboration of Zimbabwe, is projected at Batoka Gorge. Electricity distribution from Kariba extends north to the Copperbelt and southward across Zimbabwe. There are also links with Zaire and peripheral areas of Botswana and Namibia.

Agriculture, forestry, and fishing. Although contributing less than 15 percent of GDP, agriculture employs about 70 percent of the economically active population. Levels of commercialization are relatively low, and near-subsistence farming is widespread. Most agricultural produce is consumed within Zambia. A major objective of government policy is the expansion and diversification of the agricultural sector to take up the slack caused by the contraction of the mining industry. It would also contribute to lessening and perhaps reversing the rural-to-urban migration. For many years producer prices were kept low, farm incomes being depressed in favour of keeping living costs low in the towns. Increased producer prices have resulted in considerable urban unrest.

For many years the growth of maize (corn) was promoted by the use of hybrid varieties and subsidized fertilizers. It began to displace staples such as cassava, sorghum, and millet in areas not naturally suited to it (*e.g.*, higher-rainfall areas of the north). The removal of fertilizer subsidies reversed that trend; in the north, cassava, the traditional staple, is regaining importance. Where conditions are favourable and there is good access to the markets, however, improved producer prices encourage the expansion of corn cultivation. Other crops include sorghum, bulrush millet, and finger millet. Sorghum is widespread in the middle Zambezi and west of the Copperbelt. Finger millet is essentially a crop of the northeast, whereas bulrush millet is extensive in the west and along the middle Zambezi. Of the leguminous crops, groundnuts (peanuts) are most widespread, especially in the eastern part of the country and in the sandy areas of the west. Secondary food crops include sweet potatoes, taro, yams, peas, beans, pumpkins, sugarcane, bananas, rice, and a variety of other fruits and vegetables.

Cultivation methods

On the poorer soils of the wetter north and northeast, cultivation is mainly of a shifting variety called *chitemene*, whereby trees (or their branches) are cut and then piled in the centre of the clearing for burning, the crop being planted in the ashes. Over much of the rest of the country, semipermanent hoe cultivation predominates; in swamp and lakeshore areas, it is combined with fishing. Oxen are used on the sandy soils of the west.

Large-scale commercial farming has restricted distribution, mainly along the Line of Rail (notably on the Tonga plateau in the south, near Lusaka and Kabwe, on the Copperbelt, and near Mkushi) but also in the Chipata area and around Mbala at the southern end of Lake Tanganyika. At independence almost all of the 1,200 commercial farmers were of European or South African origin, but about half left in the years immediately following. Many farms were taken over by the state, but state farms have not been a success. Some were taken over, not always successfully, by industrial companies that were encouraged to invest in agriculture. The main crop is maize. Virginia tobacco has lost some of its popularity, although smaller growers have been encouraged to produce it. There has been increasing export by air of horticultural produce to European markets.

Irrigated agriculture is increasingly important. Started in 1966, the first successful scheme was at Nakambala on the south side of the Kafue Flats, where the Zambia Sugar Company has more than 25,000 acres under sugarcane. Their refinery also serves nearby smallholder cane-growing projects. Zambia provides for its own needs and exports sugar. At Mpongwe, south of Luanshya, a major irrigation scheme produces wheat and coffee. Kasama in the northeast is the location of two other arabica coffee schemes, and there is a tea estate at Kawambwa in the far north. Wheat and cotton are produced at Sinazongwe and Sinazeze in the Gwembe (middle Zambezi) valley, using water from Lake Kariba. Cotton cultivation was encouraged by the construction of textile mills, first at Kafue, later at Kabwe.

Cattle are found only in the drier, tsetse-free parts of the country with open woodland vegetation: mainly the Tonga plateau, the Kafue Flats, and the floodplain of the upper Zambezi (tsetse flies are prevalent along much of the middle Zambezi). Cattle that form part of traditional farming systems often do not enter the commercial market, which is supplied mainly by larger herds kept on commercial farms, especially near Lusaka and in the south.

Soil erosion is a perennial concern in the heavily settled areas of the south and east, while the middle Zambezi valley and the southwest are worst affected in drought years.

Forest reserves and timbering

Some 26,000 square miles of Zambia are classified as forest reserves, although the greater part of the country is wooded but not protected in this way. The main commercial timber areas are on the Copperbelt, where there have been plantings of exotic softwoods to supply the needs of the mining industry, and in the southwest, where there are extensive areas of Zambezi teak. A mill at Mulobezi, which supplies timber products, is linked to Livingstone by a light railway. A major concern is forest destruction due to charcoal burning; in the towns, charcoal is the most popular cooking fuel. The government has supported attempts to introduce energy-efficient charcoal stoves.

Zambia has relatively rich fisheries based on its many lakes, swamps, and seasonally inundated floodplains. Of particular importance is the Luapula valley, which supplies the Copperbelt. Lake Tanganyika is famous for Nile perch and kapenta, a freshwater sardine. Lusaka is supplied

mainly from the Kafue Flats and the Lukanga Swamp. Of lesser importance is the fishery on the upper Zambezi. There has been a revival of fishing on Lake Kariba, interrupted by the conflict with Rhodesia during the 1970s. There is a fishery of kapenta (a deep-feeding species caught at night using special lamps to direct their movements), which had been introduced successfully from Lake Tanganyika, although the fishery is better established in Zimbabwe, where fishing was not stopped by the war. Most fish is smoked before being trucked to market.

The Copperbelt

Industry. The Copperbelt is the country's industrial heart, the focus of mining and ancilliary industries. Local people have worked the ores for many centuries, but commercial mining essentially dates back to the 1920s. The ores occur at depth in a synclinal structure so that deep-shaft mining is normal, although there has been some opencut mining. Exhaustion of reserves and the increasing costs of mining led to the closure of the Kansanshi and Chambishi mines in the mid-1980s, and rationalization of operations in an attempt to contain costs has closed down some refining and ancilliary plants. There is much mining-related industrial activity on the Copperbelt, and a major downturn in mining activity would have severe repercussions for the area as a whole. The other major mining centre is at Kabwe, where the lead and zinc mine has been virtually exhausted. Mining elsewhere, with the exception of coal at Maambwe, is mainly small-scale.

Manufacturing industry was poorly developed before independence, most investment in this sector during the federal period being made in what is now Zimbabwe. However, during the First National Development Plan major investment was made in manufacturing, particularly import substitution. Major plants under the ZIMCO umbrella produce fertilizers, explosives, tires, hessian and grain bags, textiles, glass, cement, batteries, and foodstuffs (brewing and corn-milling being especially important). Automobiles are assembled at Livingstone.

Tourism is based mainly on game viewing and the Victoria Falls (which also offers white-water rafting in the gorges below). Although appealing to a limited market, hunting safaris are a major source of income. Tourism promotion is coordinated by the Ministry of Tourism and the National Tourist Board. Development has been handicapped by the limited number of hotel beds, poor communications, and the few alternative attractions.

Transportation. As a landlocked country, Zambia is reliant on neighbouring countries for access to the sea, and civil strife in three of these has closed key routes to the coast for much of the period since independence. At

Rail lines

that time most imports and exports were handled by the railway that linked the country with the ports of South Africa and Mozambique via Rhodesia. The 3-foot-6-inch-(1,065-millimetre-) gauge line crossed the Zambezi at the Victoria Falls in 1905, reaching the Copperbelt in 1909. To the north it was linked with the Benguela Railway in 1931, giving access to the Angolan port of Lobito.

After the Rhodesian UDI, in accordance with the sanctions policy of the United Nations, Zambia took measures to reduce its dependence on routes to the south. Much traffic was diverted to the Benguela Railway before civil war in Angola closed that route, and a project to link Zambia with the Tanzanian port of Dar es Salaam was revived. Failing to obtain Western support, the two countries turned to China for help to build the 1,060-mile Tan-Zam railway, completed in 1976. Unfortunately, the railway, which links with the older railway at Kapiri Mposhi, has not carried the projected volume of traffic, owing partly to congestion at the port of Dar es Salaam and partly to problems with track and rolling stock.

Political changes in southern Africa have lessened the need to use northern routes. South African ports are being used more, and copper can also be trucked through Namibia's Caprivi Strip and by rail from Grootfontein to Walvis Bay.

Much was done to improve the road system. The Great North Road was tarred to the Tanzanian border at Tunduma, and the Great East Road to Chipata and the Malaŵian border. Malaŵi has since extended its railway from the capital, Lilongwe, to the Zambian border at Mchinji, but Zambia has not built the 26-mile link with Chipata. Upstream of the Victoria Falls the Zambezi is unbridged, and roads on the Kalahari Sands are especially difficult.

An eight-inch petroleum pipeline from Dar es Salaam to Zambia's Indeni refinery in Ndola was completed in 1968.

Zambia's large rivers are relatively little used for transportation because of the presence of rapids and waterfalls and marked seasonal flow variations. Local transport is important on lakes. Mpulungu, a small port on the southern end of Lake Tanganyika, handles minor amounts of traffic bound for Rwanda and Burundi and links with the East African rail system.

Zambian Airways Corporation operates domestic and international services. Scheduled internal service by other operators was first allowed in 1990. The main airports are at Lusaka, Ndola, and Livingstone, but there are 12 secondary and 31 minor airports, in addition to private airstrips.

Administration and social conditions. *Government.* Zambia's initial constitution was abandoned in August 1973 when it became a one-party state. The constitution of the Second Republic provided for a "one-party participatory democracy," with the United National Independence Party (UNIP) the only legal political party. In response to mounting pressures within the country, the constitution was changed in 1991 to allow the reintroduction of a multiparty system.

Structure of the central government

Under the terms of the new constitution, the president, who is head of state and commander in chief of the armed forces, is elected by universal adult suffrage to a five-year term of office. He is empowered to appoint the vice president, the chief justice, and members of the High Court on the advice of the Judicial Services Commission. During the president's absence, his duties are assumed by the vice president. The president also appoints a cabinet from elected members of the National Assembly. The cabinet consists of 25 ministers, 30 deputy ministers (some ministries have 2), and 9 provincial deputy ministers.

The legislature, called the National Assembly, has 150 members, and elections to it are held every five years. Members of Parliament (as members of the National Assembly are usually called) are elected by universal adult suffrage. There is a 27-member House of Chiefs, with a two-year-term rotating membership. It has no legislative function: it may consider bills but not block their passage.

Central government is represented throughout Zambia by the provincial government system, by which resident ministers are appointed by the president to each of the nine provinces (including Lusaka Province, created 1976). Each minister is the president's direct representative and responsible for the coordination of policy and for liaison with district councils.

The nine provinces are divided into 55 districts, each of which has a district council chairman responsible to the provincial deputy minister; the district council chairman is particularly concerned with political and economic developments. His civil service counterpart is the district executive secretary. The cities of Lusaka, Ndola, and Kitwe have councils and mayors, but the formerly separate management of mine townships on the Copperbelt has been abolished.

Justice. The court system consists of the Supreme Court, the High Court, subordinate magistrate's courts, and local courts. Because the law administered by all except the local courts is based on English common law, decisions of the higher British courts are of persuasive value; in fact, a few statutes of the British Parliament that were declared by ordinance (decree) to apply to Zambia are in force so far as circumstances permit. Most of the laws presently on the statute book, however, have been locally enacted by ordinance or, since independence, by Zambian acts.

The Supreme Court consists of the chief justice and four other justices; it is the court of last resort. The High Court, presided over by a chief justice, has 12 puisne judges and is basically an appellate court. There are three classes of magistrate's courts, with progressive degrees of criminal and civil jurisdiction. Local courts consist of a president sitting alone or with other members, all appointed by the

Judicial Services Commission. Jurisdiction is conferred by the minister of justice and may encompass any written law, but punishment powers are limited. Local courts also deal with civil cases of a customary nature. Customary law is followed when it is not repugnant to justice or equity and when it is not incompatible with other legislation.

Independence of the judiciary

The judiciary remains formally independent, and in this respect Zambia contrasts favourably with many other African countries. The president appoints the chief justice and, on the advice of the Judicial Services Commission, also appoints other judges; however, the constitution severely restricts the president's powers of dismissal, and on occasion judges have not shrunk from challenging the authority of the government or party. At the same time, the scope of the judiciary was seriously limited by presidential powers of preventive detention under emergency regulations brought in at the time of Rhodesian UDI in November 1965 and subsequently regularly renewed by the National Assembly. The ending of these state-of-emergency regulations on Nov. 8, 1991, was one of the first acts of the new government.

Education. At independence Zambia had one of the most poorly developed education systems of Britain's former colonies, with just 109 university graduates and less than 0.5 percent of the population estimated to have completed primary education. The country has since invested heavily in education at all levels, and well over 90 percent of children in the 7–13 age group attend school. However, of those who enroll for the seven years of primary education, less than 20 percent enter secondary school, and only 2 percent of the 20–24 age group enter university or some other form of higher education.

The University of Zambia was opened in Lusaka in 1966, graduating its first students in 1969. In 1979 legislation was passed creating a federal university; a second campus was established at the Zambia Institute of Technology at Kitwe. In 1988 the federal structure was abandoned, and Zambia now has two universities: the University of Zambia at Lusaka and the Copperbelt University at Kitwe. The former offers courses in agriculture, education, engineering, humanities and social sciences, law, medicine, mining, natural sciences, and veterinary medicine, but only business and industrial studies and environmental studies are available at Kitwe. The basic program is four years, although engineering and medical courses are of five and seven years' duration, respectively.

Other tertiary-level institutions are vocationally focused and include the Evelyn Hone College of Applied Arts and Commerce and the Natural Resources Development College, both in Lusaka, as well as teacher-training colleges.

Early developments in continuing education have been undermined owing to underfunding. A 1976 initiative to introduce a major reform of the educational system was thwarted by the economic downturn, and underfunding is lowering the quality of education.

Effects of AIDS

Health and welfare. The ailing economy in the 1980s adversely affected the quality of health care available to the population at the time that AIDS (acquired immune deficiency syndrome) was beginning to have a major impact. Zambia is one of the southern African countries most severely affected by AIDS, with an estimated 20 percent of the urban population HIV-positive. Early deaths from AIDS-related illnesses are depriving the country of expensively trained skilled professionals and creating a growing number of orphaned children. Malnutrition, caused by poverty, is widespread, particularly in the rural areas, and is a major cause of death among children. The most prevalent tropical diseases are malaria, schistosomiasis (bilharziasis), and parasitic infections such as hookworm and leprosy. Leprosy has been contained, and leprosariums have given way to outpatient treatment. Malaria is increasing in the urban areas as programs to control the anopheles mosquito that spreads the disease have largely broken down. Schistosomiasis, a debilitating disease spread by waterborne snails, is widely found in riverine areas. Sleeping sickness, spread by the tsetse fly, is prevalent in the more sparsely populated tsetse-infected areas. Smallpox and typhoid fever have been successfully controlled through immunization programs. By contrast, there have been major outbreaks of cholera and dysentery in Lusaka and the Copperbelt, undoubtedly associated with increasing poverty and deficiencies in sanitation and community health programs. Blindness is a particular problem in the Luapula valley.

Tuberculosis and meningitis, related to AIDS, are now major causes of adult and infant mortality. Other common causes of death are respiratory infections, accidents and injuries (relative to the number of vehicles, the number of motor vehicle accidents is exceptionally high), and gastrointestinal disorders. Measles is a common cause of death in children. Death from heart disease is rising among the more affluent.

In the years following independence, considerable investment was made in the hospital system, which includes 12 general hospitals in the main towns, many smaller hospitals (some of which are mission-run), and rural health centres. The University Teaching Hospital in Lusaka is used by the medical school of the University of Zambia, which graduated its first doctors in 1972. A government Flying Doctor Service provides medical services in remote rural areas. Psychiatric services are based at the Chainama Hills Hospital in Lusaka, to which are linked small psychiatric units in other centres. There is a specialist pediatric hospital in Ndola. Despite local training, Zambia suffers from a shortage of doctors and other specialist staff. This is particularly true of the rural areas, although there are some excellent mission hospitals. There is a widespread belief in alternative medicine, including reliance upon traditional healers and witchcraft.

Social services

In traditional Zambian society, kinship groups care for the well-being of their members. In the towns, however, family ties have weakened, necessitating the development of government welfare services concerned with juvenile delinquency, adoption, and the care of the aged, indigent, and handicapped. Voluntary agencies, many drawing upon funds from outside Zambia, make a major contribution to the care of the less fortunate in society. The contributory National Provident Fund provides retirement benefits for those in paid employment (a minority of the labour force, including many town dwellers, are engaged in informal employment). Refugees have been a major problem, notably those fleeing conflicts in Angola and Mozambique, and the country once gave haven to many who had fled from Rhodesia during UDI and from South Africa.

Housing is a problem in the urban areas owing to the high level of rural-to-urban migration. Public housing could be made available only to a few, and shanty compounds sprang up to house the majority. "Site and service schemes" designate areas for self-help housing and provide basic services such as roads and water. In Lusaka the World Bank assisted with major schemes to upgrade existing squatter areas. Nevertheless, there is a sharp contrast between the spacious bungalows in the leafy suburbs, many built for Europeans before independence but now occupied by wealthy Zambians, and the cement-block and tin-roofed houses of the dusty and crowded townships.

Cultural life. Traditional Zambian art consists chiefly of wood carving, pottery making, and basket weaving. Among musical instruments, drums are the most widely used, but there also are stringed bows, flutes, horns and pipes, xylophones, bells, rattles, and the kalimba, or "African piano," made of strips of steel attached to a small board and vibrated by the fingers. Music, dancing, and song are used in tribal rituals and celebrations, as well as for entertainment, varying in form among ethnic groups. With the object of preserving cultural diversity, a government initiative in the 1980s led to the revival of many traditional ceremonies. Some, such as the *kuomboka* of the Lozi, survived essentially unchanged; others have taken up new forms. The National Dance Troupe performs the traditional dances of many groups. There is a national museum at Livingstone and another on the Copperbelt. The Moto-Moto Museum at Mbala focuses on the traditions of the Bemba people, and there are small field museums at some national monuments. Relics of the country's past are the concern of the Commission for the Preservation of Natural and Historical Monuments and Relics.

Since the 1950s the cultural scene has been transformed

by large-scale urbanization and exposure to exotic influences from Europe, the Americas, and other parts of Africa. Radio and television (one channel only and restricted to the Line of Rail, Chipata, Mongu, and Kasama) have sped this process. Various forms of theatre have flourished. In the last years of colonial rule, dance drama was developed for nationalist ends; the Chikwakwa Theatre, based at the University of Zambia, pioneered politically radical popular drama in the early years of independence. In the 1980s, aid agencies and other bodies promoted "theatre for development," often unscripted and in vernacular languages, and government departments have used drama to communicate agricultural and health messages.

Publishing

The once-noted independence of the press was compromised by government and party control: ownership of the *Times of Zambia* passed to the party in 1982, and the *Zambia Daily Mail* has been in government ownership since 1965. The church-owned weekly *National Mirror,* founded in 1970, was able to take a more independent line, as has the *Weekly Post,* which first appeared in 1991 in the wake of moves toward a more pluralist political system.

The Zambia Publishing House (formerly the Kenneth Kaunda Foundation) is a government-backed publisher of the works of Zambian authors and school textbooks. The few other publishers are mainly church-supported. Zambian scholars have contributed to knowledge in a wide range of disciplines, often in locally published academic journals, though opportunities for research have been restricted in recent years by general economic difficulties.

(Ri.H./G.J.Wi.)

For statistical data on the land and people of Zambia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Archaeology and early history. Stone tools attributable to early types of man have been found near the Victoria Falls and in the far northeast, near Kalambo Falls. In 1921, excavations at Kabwe revealed the almost complete skull of *Homo sapiens rhodesiensis* ("Broken Hill Man"), which may be well over 100,000 years old. However, by 20,000 BC the only surviving type of human throughout the Old World was the ancestor of modern man, *Homo sapiens sapiens,* who developed the use of spears, the bow and arrow, game traps, and grindstones. Remains of such industries have been found in much of central and northern Zambia, sometimes near lakes and rivers but often in caves and rock-shelters.

Bantu migrants

During the 1st millennium AD, Zambia was occupied by migrants from farther north who may be presumed to have spoken Bantu languages; they certainly cultivated crops and kept domestic stock. Traces of ironworking in central and western Zambia have been dated to the first five or six centuries AD. Iron tools and weapons greatly increased mastery over both man and nature and, together with food production, promoted population growth. Stone-using hunters and gatherers were liable to be overrun and absorbed by the food producers, though some survived on the edges of farming zones until a few centuries ago. The complex layers of paintings found in rock-shelters in northeastern Zambia indicate that the homes of stone-using hunters became the shrines of invading farmers.

In central Zambia, by the 6th century AD, the first food producers worked copper as well as iron. By about AD 1000, copper ingots were being made at Kansanshi, at the western end of the Copperbelt, which implies that copper was being traded extensively and perhaps used as currency.

Early in the 2nd millennium AD, cattle keeping became more intensive on the Batoka Plateau of southern Zambia, while cotton spinning and pipe smoking were introduced. The associated pottery seems directly ancestral to that made locally in the 20th century. Similar evidence of cultural continuity over a long period has also been found in the resemblance between modern pottery in central, northern, and eastern Zambia and a kind of pottery that has been dated to the 12th century AD. These major changes in pottery traditions have been ascribed to immigration; they also indicate thicker settlement of woodland through the adoption of *chitemene* cultivation, widespread in Zambia even today: this technique depends heavily on the use of iron axes, because seed is sown in the ashes of branches lopped from trees.

In southern Zambia, archaeology has thrown light on both the emergence of class distinctions and the beginnings of trade with the east coast. About the 14th century, a few people were buried wearing ornaments of seashells and exotic glass beads near Kalomo and at Ingombe Ilede, near the confluence of the Zambezi and Kafue. The latter burials also included gold beads, copper ingots, and iron bells of a kind later associated with chieftainship. These metals would have come from south of the Zambezi, but they were probably being reexported down the river by Muslim traders, either Arab or African.

The period between about 1500 and 1800 remains relatively obscure. This was when copper was most intensively mined at Kansanshi, but it is not known who was buying it. The main evidence for these centuries consists of oral traditions. In much of Zambia, from the upper Kafue to the Malaŵi border, there are legends of tribes being founded by chiefly families who came from Luba country in southeastern Zaire. Such stories should not be taken at face value: they dramatize prolonged processes of population drift and the spread of cultural influences. By the 18th century, small-scale chieftainship was probably widespread in northern and eastern Zambia, but few of the tribal names current today would have meant much: such names refer not to long-enduring communities but to changing perceptions of cultural and political differences.

Early 19th-century kingdoms

In the early 19th century, however, there were at least four areas in which the growth of kingdoms was strengthening the sense of tribal identity: in the east, among the Chewa; in the northeast, among the Bemba; on the lower Luapula, among the Lunda (who had indeed invaded from the west about 1740); and on the upper Zambezi, among the Luyana (later called Lozi). In the Lunda and Luyana kingdoms a prosperous valley environment encouraged dense settlement and prompted the development of relatively centralized government.

External contacts. Trade between Zambia and the Western world began with the Portuguese in Mozambique. Early in the 17th century, the Portuguese ousted Muslims from the gold trade of central Africa; early in the 18th century, they founded trading posts at Zumbo and Feira, at the confluence of the Zambezi and Luangwa; and by 1762 they were regularly acquiring ivory and copper from Zambians in exchange for cotton cloth. During the later 18th century, slave-owning Goans and Portuguese mined gold and hunted elephants among the southern Chewa. Their activities were reported to Kazembe III, the Lunda king on the Luapula, by Bisa traders who exported his ivory and copper to the Yao in Malaŵi. Kazembe already had indirect access to European goods from the west coast; he now hoped to cut out his African middlemen. One Goan visited Kazembe and was warmly received, but, though the Portuguese government dispatched further expeditions in 1798 and 1831, they came to nothing, mainly because the Portuguese on the Zambezi were turning their attention to exporting slaves rather than ivory or gold. Western Zambia was also beginning to be enmeshed in the Portuguese slave trade (directed to Brazil): from the early 19th century, African traders from Angola bought slaves to the north of the Lozi kingdom, though the Lozi themselves kept servile labour for production at home.

During the second half of the 19th century, Zambia was convulsed by traders, raiders, and invaders who came from north and south as well as east and west. From about 1840 to 1864, the Lozi kingdom was ruled by the Kololo, warrior-herdsmen who had fled north from Sotho country. In the 1860s and '70s the northern Chewa were conquered by a group of Ngoni, who had also come from the far south. Meanwhile, the Bemba and Kazembe's Lunda began selling ivory and slaves to Arabs and Africans from the east coast. At the same time, ivory and slaves were hunted in central Zambia by Chikunda adventurers armed with guns, and South African traders were buying ivory from the Lozi. A few rulers contrived to turn such trade to their own advantage, and the general

rise in demand for goods stimulated local production of ironwork, salt, tobacco, and food: indeed, several crops of American origin were introduced, such as corn (maize), cassava, peanuts, and sugarcane. Much of Zambia was devastated by marauders, however.

British rule

At the end of the 19th century, Zambia came under British rule. British interest in the region had first been aroused by the missionary-explorer David Livingstone, who crossed Zambia during three great expeditions between 1853 and his death, near Lake Bangweulu, in 1873. Livingstone's reports of the expanding slave trade inspired other missionaries to come to central Africa and continue the struggle against it, but it was the mining magnate Cecil Rhodes who ensured that so much country north as well as south of the Zambezi came within a British sphere of influence during the "Scramble for Africa." In 1889 the British government granted a charter to Rhodes's British South Africa Company (BSAC), bestowing powers of administration and enabling it to stake claims to African territory at the expense of other European powers. The curious butterfly shape of Zambia resulted from agreements in the 1890s between Britain and Germany, Portugal, and the Belgian king Leopold II, and these in turn rested on treaties, mostly stereotyped in form, between Rhodes's agents and African chiefs.

At this stage there was little resistance to white intrusion. The most immediate threat to African land and labour came in Ngoniland, thought by whites to be rich in gold, and the Ngoni duly fought company troops in 1898. The Bemba, however, faced no such challenge and in any case were deeply divided, while the Lozi king believed that alliance with the company would protect his empire against both the Portuguese and the Ndebele. It is also likely that disease and famine undermined the will to resist: there were smallpox epidemics in the early 1890s, widespread rinderpest in 1892–95, and locust plagues throughout the decade.

Formation of the Rhodesias

Colonial rule. At first the BSAC administered its territory north of the Zambezi in two parts, North-Eastern and North-Western Rhodesia. In 1911 these were united to form Northern Rhodesia, with its capital at Livingstone, near the Victoria Falls. The population was probably about one million. There were about 1,500 whites: some came to mine surface deposits of copper, and a few, mostly from South Africa, farmed on the plateau east of Livingstone. However, the BSAC regarded the country chiefly as a source of labour for gold and coal mines in Southern Rhodesia and for the copper mines in Katanga, in the Belgian Congo, which in 1910 were linked by rail to Southern Rhodesia and the east-coast port of Beira, Mozambique. By then, company officials had been posted to most parts of Northern Rhodesia and levied taxes in order to force Africans to seek work: such pressure sometimes provoked violent, but small-scale, resistance.

World War I bore heavily on the territory. For the campaign against the Germans in East Africa, 3,500 troops were recruited and 50,000 porters conscripted, mostly from the northeast; many never returned. Food supplies were requisitioned, yet food production was crippled; women, as always, bore the brunt of sowing and harvesting, but, in the absence of men to cut trees and clear new land, farm plots were worked to exhaustion. Labour was also urgently needed for mining: war boosted the demand for base metals from Northern Rhodesia as well as Katanga. The Bwana Mkubwa mine exported copper from 1916 to 1918, and from 1917 to 1925 the country's main export was lead from Broken Hill (now Kabwe). African resentment of wartime hardship found expression in the millennial Watchtower movement, which inspired rebellion among the Mambwe in the northeast. More effective opposition to BSAC rule came from white settlers, especially when an income tax was imposed in 1920. The company was ready to give up the increasingly costly burden of administering Northern Rhodesia and in 1924 handed over this responsibility to the Colonial Office in London, which soon set up a legislative council to which five members were elected by the white population, then about 4,000.

The British government hoped to increase white settlement as part of a wider strategy to strengthen British influence between South Africa and Kenya. Land was reserved for white ownership along the railway line, in the far north, and in the east; around these areas African reserves were marked out in 1928–30. This soon led to overcrowding, soil exhaustion, and food shortage; yet few whites took up the land available to them. By 1930 it was clear that copper was the country's most promising resource. Huge deposits had been located far beneath the headwaters of the Kafue and were mined by companies mostly financed from South Africa, through the Anglo American Corporation, and the United States, through the Rhodesian Selection Trust.

The mining industry

In 1930–31 prices for copper collapsed, partly as a result of the worldwide depression. However, the new mines enjoyed a comparative advantage, since they worked high-grade ores at relatively low cost. For skilled labour, they depended on whites who had to be paid what they might have earned in South Africa; but African labour was cheap and abundant, and employers accepted a high turnover rate rather than provide the amenities that would encourage permanent African settlement in urban areas. From 1935, copper prices rose sharply, and by 1938 Northern Rhodesia contributed more than 13 percent of noncommunist copper production.

Yet copper exports did not confer much prosperity. Near the railway, African as well as white farmers grew food for the mines, but most African farmers were too remote from the market to be able to earn a cash income. More than half the able-bodied male population worked for wages away from home, and as many of these worked outside the territory as within it. On the Copperbelt itself, low wages and poor conditions provoked Africans to strike at three mines in 1935. Nor were rising copper sales of much benefit to the government (whose capital was moved to Lusaka in 1935). The mineral rights were owned by the BSAC, which duly exacted royalties. Taxation was levied on what profits remained, but half was retained by the British government, which made only tiny grants for economic development. In 1938 these arrangements were criticized by a visiting financial expert, Sir Alan Pim: in a report to the Colonial Office, he urged more public investment in roads, schools, and health services, for Africans as well as whites. Missionaries ran many primary schools, but in 1942 only 35 Africans were receiving secondary education.

When World War II broke out in 1939, Britain contracted to buy the whole output of the Copperbelt. British dependence on undisturbed copper production meant that white mine workers were allowed to maintain an industrial colour bar. Nonetheless, a second strike by African mine workers, in 1940, caused a revision of wage scales to take account of accumulating experience and skill. After the war the new Labour government in Britain began to promote the formation of African trade unions, and by 1949 half the African mine workers in Northern Rhodesia belonged to a single union. In the same year, new legislation confirmed that (in contrast to South Africa and Southern Rhodesia) African unions had the same bargaining rights as those of white workers. Meanwhile, between 1942 and 1946, African teachers, clerks, foremen, and clergy had formed welfare societies both in the mining towns and in rural areas; in 1948 these gave rise to the Northern Rhodesia Congress. Some of its members sat on the African Representative Council set up by the government in 1946; this body had no power, but it criticized political and social conditions, especially the informal colour bar, and from 1948 it elected two Africans to sit on the Legislative Council. In the countryside, "indirect rule" through chiefs became more broadly representative.

In some respects, Africans made important advances in the first postwar years. On the other hand, these advances also strengthened white aspirations to settler self-government, as in Southern Rhodesia. Although whites formed less than 2 percent of the Northern Rhodesian population, their numbers rose between 1946 and 1951 from 22,000 to 37,000, partly because of immigration from Britain. The Legislative Council included eight elected white members, and in deference to them a large-scale development plan was drastically revised between 1947 and 1953 at the expense of African education. Yet this was not enough:

to many whites the best hope of entrenching white supremacy seemed to lie in amalgamation with the south. This ambition gained support from British politicians and civil servants who feared that Southern Rhodesia would otherwise fall under the sway of the Afrikaner nationalists who had come to power in South Africa in 1948. In 1951 the British Labour government was replaced by Conservatives less concerned to avoid alienating African opinion. Despite widespread popular protest, in which chiefs and Congress combined, in 1953 Northern and Southern Rhodesia and Nyasaland were brought together in the Central African Federation.

Creation of the Central African Federation

The federation was a curious and unstable compromise. Its government was based in Southern Rhodesia, which also dominated the federal parliament. It had wide powers over all three territories, though in the north Britain retained control over questions of African land, education, and political status. At first, African suspicions of federation were blunted in Northern Rhodesia by an economic boom. Copper prices had risen steeply following sterling devaluation in 1949 and the outbreak of war in Korea in 1950; the mining companies finally began to pay regular dividends, while the Northern Rhodesia government received a share of royalties. Following a major African strike in 1952, the real wages of African mine workers at last moved upward. The companies increased their use of machinery and African skills; in 1955 the industrial colour bar was breached, and a select minority of African workers were encouraged to live out their working lives in the mining areas: "stabilized" labour began to replace oscillating migrant labour.

In 1956, however, the copper boom came to an end. Whites in Northern Rhodesia became increasingly aware of how far the federal tax system channeled copper profits into Southern Rhodesia. Many Africans were thrown out of work, while little had been done to help African farming or education, despite federal propaganda for "partnership." A new generation of Congress leaders wanted Northern Rhodesia to become an independent African state, as Ghana had in 1957. In 1958, led by Kenneth Kaunda, a former teacher and civil servant, these radicals split off from Congress to found the Zambia African National Congress and its successor, the United National Independence Party (UNIP). Britain accepted that Africans would have to be given more power than the federal government was willing to concede. In 1962 UNIP organized a massive campaign of civil disobedience, but it agreed to take part in elections under a new constitution: an election later that year gave Africans a majority in the legislature. The federation was dissolved at the end of 1963. Early in 1964 an election based on universal adult suffrage gave UNIP a decisive majority, and it was supported by nearly a third of the white voters. On October 24 the country became the independent Republic of Zambia, within the Commonwealth and with Kaunda as executive president.

Formation of UNIP

Zambia since independence. During the early years of independence, Zambia was comparatively prosperous. Copper prices rose steadily from 1964 to 1970, boosted by the Vietnam War, and Zambia became the world's third largest producer of copper. Meanwhile, the leakage of copper profits abroad was greatly reduced. In 1964 the government acquired the mineral rights of the BSAC, and thereafter it also increased mining taxation. The country embarked on long-overdue investment in communications and social services. In 1960 there had been only 2,500 Africans in secondary schools; by 1971 there were 54,000. At independence there were fewer than 100 Zambian university graduates; in 1965 the University of Zambia was founded, and by 1971 it had 2,000 students. Zambians finally began to predominate in the upper ranks of the civil service, the army, business, and the professions. The copper industry still relied heavily on white expertise, but the colour bar had vanished, and in 1966 black mine workers secured a large increase in pay, which soon affected wage levels generally.

On the other hand, Zambia incurred massive costs from the survival of white supremacy across the Zambezi. Following (Southern) Rhodesia's Unilateral Declaration of Independence (UDI) in 1965, the United Nations imposed sanctions intended to isolate that country, but these bore much more heavily on Zambia. Copper exports were expensively rerouted northward, and a tarmac road and oil pipeline were built to Dar es Salaam, Tanz. Trade with Rhodesia was steadily reduced, and the border was finally closed in 1973. A new coal mine and new hydroelectric schemes made Zambia largely independent of the Rhodesian-controlled power station at the Kariba Dam (built in 1959). In 1970–75 China built a railway from the Copperbelt to Dar es Salaam, which committed Zambia and Tanzania to extensive trade with China and to a huge loan for which payment was deferred until the 1980s.

National integration had been a major task for Zambia's leaders at independence. White settlers presented no great difficulty, and those farmers who stayed on were valued for their major contribution to food production. African "tribalism" was a more serious problem. This had less to do with the survival of precolonial political loyalties than with regional differences aggravated under colonial rule and the absence of any African lingua franca. The Lozi and other peoples in the west and south had long depended on labour migration across the Zambezi; the Copperbelt was dominated by Bemba speakers from the northeast. Kaunda did not himself belong to any major ethnic group, but his continuation in power required constant reshuffling of colleagues in the party and the government to preclude the emergence of a rival. In the name of national unity, UNIP sometimes made exaggerated claims to allegiance; such claims had brought it into armed conflict in 1964 with the Lumpa church founded by Alice Lenshina and in the late 1960s with Jehovah's Witnesses. UNIP also challenged the independence of the judiciary, though from 1969 the authority of the bench was strengthened by the appointment of black Zambian judges.

National integration

In the early 1970s Zambia's economic fortunes took a turn for the worse. Copper continued to provide the great bulk of export earnings, but prices fluctuated erratically; they then failed to recover from a steep fall in 1975. The price of oil shot up in 1973, and inflation, already serious, rapidly increased. The government, committed to high spending, both public and private, reacted by borrowing heavily abroad and drawing on reserves. Investment declined, as did the efficiency of the transport network. State control of the mining industry, achieved in 1969–75, artificially prolonged its life but also increased the scope of corruption, as did parastatal corporations set up to promote industrial diversification.

The government became increasingly authoritarian. Kaunda felt threatened by critics at home and by the illegal Rhodesian regime, which harassed African guerrillas based in Zambia. UDI had already prompted Kaunda to impose emergency regulations, which thereafter were regularly renewed by Parliament and enabled the president to detain political opponents without trial. In 1973 Parliament approved a one-party constitution, and in 1975 UNIP took over Zambia's main newspaper. To some extent, fear of foreign attack diminished with the advent of independence in Portuguese Africa in 1975 and in Rhodesia (Zimbabwe) in 1980. But warfare in Angola and South African interference continued to provide pretexts to curb internal opposition.

The move to a one-party state

Still more worrying, however, was the deepening economic crisis. Kaunda urged Zambians to look to agriculture rather than mining for a solution, but rural development policies, though consuming foreign aid, were mostly ill-conceived and failed to stem the historic drift to the towns. By 1980, out of a population of 5.7 million, more than 2 million lived in towns, many without jobs or housing: inevitably, disease and crime flourished. Urban dwellers refused to pay the high prices that might have encouraged more farmers to produce for the market. Government subsidies sometimes bridged the gap, but their partial removal in 1986 and 1990 provoked major food riots in the towns. The restoration of subsidies in 1987 cost Zambia the support of the International Monetary Fund, though such support had been critical in coping with enormous foreign debts. Mounting discontent was reflected in recurrent closings of the university, and in August 1991, in response to widespread pressure, the National Assembly

abolished the one-party state. Multiparty elections were held in October, and Kaunda was decisively defeated by a trade union leader, Frederick Chiluba, who duly became president. UNIP was left with fewer than one-sixth of the seats in the National Assembly.

Zambia's experience in the 19th and 20th centuries has in many ways been typical of the Third World. External trade has been primarily based on extractive forms of production: the export of ivory, labour, and copper. Profits from copper eventually made possible widespread cultural as well as economic involvement with the developed world, but by then that world was losing interest in what Zambia had to offer. Expanding state control and foreign loans provided cushions against adverse trading conditions but masked systemic inefficiency and could not indefinitely postpone a day of reckoning. Throughout the 20th century, Zambia's difficulties were indeed compounded by its proximity to white supremacies farther south, but this very fact gave other powers an interest in maintaining political stability within Zambia after independence: in effect, loans and aid helped to dampen internal economic conflict. In 1991 the prospect of fundamental change in South Africa raised the question of whether Zambia would continue to attract such interest or would be relegated to the margins of international concern. (A.D.R.)

For later developments in the history of Zambia, see the BRITANNICA BOOK OF THE YEAR.

Zimbabwe

Zimbabwe is a republic in southern Africa. It achieved majority rule and internationally recognized independence in April 1980 following a long period of colonial rule and a 15-year period of white-dominated minority rule, instituted after the minority regime's so-called Unilateral Declaration of Independence (UDI) in 1965.

Zimbabwe shares a 125-mile (200-kilometre) border on the south with the Republic of South Africa and is bounded on the southwest and west by Botswana, on the north by Zambia, and on the northeast and east by Mozambique. Its total area is 150,873 square miles (390,759 square kilometres). The capital is Harare (formerly called Salisbury).

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Zimbabwe lies almost entirely over 1,000 feet (300 metres) above sea level. Its principal physical feature is the broad ridge running 400 miles from southwest to northeast across the entire country, from Plumtree near the Botswana frontier through Gweru (formerly Gwelo) and Marondera (formerly Marandellas) to the Inyanga Mountains, which separate Zimbabwe from Mozambique. About 50 miles wide, this ridge ranges in altitude from 4,000 to 5,000 feet, until it eventually rises to 8,504 feet (2,592 metres) at Mount Inyangani, the highest point in Zimbabwe, in the eastern highlands. This ridge is known as the Highveld and comprises about 25 percent of the country's total area. On each side of this central spine, sloping down northward to the Zambezi River and southward to the Limpopo River, lies the wider plateau of the Middleveld, which, at an altitude between about 3,000 and 4,000 feet, makes up roughly 40 percent of Zimbabwe's area. Beyond this again and mostly in the south, where the Sabi, Lundi, and Nuanetsi rivers drain from the plateau into the Limpopo, lies the Lowveld, which constitutes approximately 23 percent of the country's total area. The lowest point in Zimbabwe lies at an altitude of 660 feet near Dumela, where the Limpopo flows into Mozambique. There are no parts of Zimbabwe that can properly be called desert, although a sector northwest of Plumtree and a lengthy belt across the Lowveld in the south are severely arid.

Geology

The landscape is characterized by extensive outcroppings of Precambrian rock, which is between 570 million and 3.8 billion years old. The most ancient part of this rock formation, known as the basement complex, covers the greater part of the country. About four-fifths of the basement complex consists of granite; the Matopo (Matopos) Hills south of the city of Bulawayo are formed from prolonged erosion of an exposed granite batholith. Some of the hills are surmounted by formations, known as balancing rocks, that have been eroded by wind and water along regular fault lines, leaving some blocks precariously balanced upon others. Elsewhere are found innumerable small rounded granite hillocks known locally as kopjes. Belts of schist in the basement complex contain the veins and lodes of most of the country's gold, silver, and other commercial minerals.

The Great Dyke, which is up to 8 miles wide and about 330 miles long, is another notable landscape feature. The longest linear mass of mafic and ultramafic rocks in the world, the Great Dyke bisects the country from north to south and contains enormous reserves of chromium, nickel, and platinum. The Alkali Ring complexes near Beitbridge in the Sabi valley are distinctive igneous intrusions. The Karroo (Karoo) System—a thick layer of sedimentary rocks consisting of shale, sandstone, and grit of Permian and Triassic age (208 to 286 million years old)—covers the Zambezi valley and the valleys of its tributaries from Hwange (formerly Wankie) southward to Bulawayo and spreads across parts of the southern Lowveld from Tuli, near the southern border, to the Sabi River.

Drainage and soils. Major faulting from southwest to northeast formed the middle Zambezi trough, which is now partially flooded by the Lake Kariba reservoir. Other faulting episodes affected the depressions of the Sabi and Limpopo rivers. Except for a small area of internal drainage in the dry southwest, these three rivers carry the entire runoff of the country to the Indian Ocean via Mozambique. The central ridgeline of the Highveld is the major divide separating Zambezi from Limpopo-Sabi drainage.

The light, sandy soils found in most parts of Zimbabwe are residual soils developed largely from the granite parent material. They are highly weathered and leached, even in the areas of lower rainfall, and do not easily retain water because of their coarse texture. Outcrops of basement schists give rise to rich red clays and loams—some of the country's best soils—but their extent is limited. Since most rain occurs in heavy showers during a few months of the year, rapid runoff and high rates of erosion are common. The meagre mineral reserves in most soils imply an inherently low fertility; under cultivation, productivity drops rapidly after a few years. The difficulty of cultivating these lighter soils is greatest in the black farming areas, where population pressure no longer allows land to be temporarily abandoned to rejuvenate after cultivation; black farmers, because of a lack of capital, are also less able than white farmers to maintain the mineral fertility with manure and chemical fertilizers.

Low fertility of soils

Climate. Zimbabwe, lying north of the Tropic of Capricorn, is completely within the tropics but enjoys subtropical conditions because of its high average elevation. Toward the end of the hot, dry months, which last from August to October, monsoon winds that have crossed the Indian Ocean and Mozambique result in intense orographic rainfall when they meet the rampart formed by the eastern highlands. The eastern regions consequently receive the country's heaviest rainfall and have a more prolonged rainy season (lasting from October into April) than the rest of Zimbabwe. The high altitude of the broad plateau of western Zimbabwe helps to guarantee fine weather there during the cool, dry winter months from May to August.

June is generally the coolest month and October the warmest; temperature variations correspond closely to altitude. Inyanga, at about 5,500 feet in the eastern highlands, varies in temperature from a mean of 52° F (11° C) in July to one of 65° F (18° C) in October. Harare, at about 4,800 feet, has seasonal temperatures varying from 57° F (14° C) to 70° F (21° C), and Bulawayo, at 4,400 feet, varies from 57° F (14° C) to 70° F (21° C). Daily variations about these means are some 13° F (7° C) warmer in the afternoon and 13° F (7° C) cooler at night. Harare and Bulawayo each average about eight hours of sunshine per day, and this average does not drop below six hours during the rainy season.

Plant and animal life. Zimbabwe is predominantly savanna (tropical grassland), with a generous tree growth

encouraged by the wet summers. The only true forests, however, are the evergreen forests of the eastern border and the savanna woodland, which includes teak, northwest of Bulawayo. Various species of *Brachystegia* (a hardwood tree up to 90 feet high with pale reddish brown wood) are dominant in the Middleveld and Highveld. Other common species include the mohobohobo (a medium-size tree with large spadelike leaves) and the thorn tree. In the valleys of the Zambezi and Limpopo rivers, the mopane, which resembles the mohobohobo, is common, together with the stout-trunked baobab and the knobby thorn tree. Australasian eucalyptus trees have been widely introduced, predominantly on white-owned farms, where they are used as windbreaks and for fuel; Australian wattle has been planted in the eastern districts as a source of tannin. Pure grassland is uncommon but occurs particularly along the eastern border around Chimanimani (Mandidzudzure, formerly Melsetter).

Cultivation of the land and the reduction of the natural vegetation have resulted in the disappearance of many forms of animal life over large areas. Hwange National Park, holding some of the densest remaining wildlife concentrations in Africa, has an area of more than 5,000 square miles and stretches from the Bulawayo–Victoria Falls railway line westward to the Botswana border. Among the flesh-eating animals found there, and occasionally elsewhere, are the lion, leopard, cheetah, serval, civet, aardvark, spotted and brown hyena, black-backed and side-striped jackal, zorille, ratel, bat-eared fox, ant bear, and scaly anteater. Elephants are found in the northern region, and giraffes in the western bushland; hippopotamuses and crocodiles live in the larger rivers. Among a great variety of hoofed and horned ruminant animals are the eland (which is immune to the deadly tsetse fly), greater kudu, blue duiker, impala, klipspringer, steenbok and grysbok, and sable and roan antelope. Snakes include mambas, boomslangs, and the black-necked cobra. Baboons, which are the bane of farmers whose crops they damage, include the Rhodesian and yellow species, as well as the chacma, the largest known baboon species. Notable among the birdlife are the martial eagle, the bateleur eagle, and the little hammerhead, which builds enormous nests and is revered as a bird of omen.

Hwange National Park

Settlement patterns. Zimbabwe may be divided into six different regions of agricultural potential, with the amount of rainfall constituting the determining factor in land use. The eastern highlands, with more than 25 inches of rainfall annually, are suitable for diversified farming with cattle and plantation and orchard crops. Roughly one-fifth of the country, sweeping west along the central spine past Harare and on to the midlands, receives 20 to 25 inches of rain and is used for intensive farming of corn (maize) and tobacco and the raising of livestock. An almost equal area to the southwest, enclosing Bulawayo, receives 16 to 20 inches of rain a year; it is suitable for mixed farming and for raising livestock on a semi-intensive scale. One-third of the country, lying farther outward from the spine of Zimbabwe, mostly to the south, and receiving 14 to 18 inches of rainfall annually, is used for semi-extensive farming, while about one-fourth of the country in the Lowveld toward the Limpopo and Zambezi rivers, receiving less than 16 inches a year, is fit only for ranching. Finally, a small area, mostly in the far north toward the Zambezi River, is unsuitable for either agriculture or forestry.

Land ownership

Prior to independence most of the country's best farmland was in the hands of white settlers or absentee landlords. In consequence, the nationalist struggle focused sharply upon the issue of land ownership, and a major concern for the Zimbabwe government after independence was to carry through land reform in the rural areas and launch large-scale settlement of black families on former white farms.

The Land Apportionment Act, a segregationist measure that governed land allocation and acquisition prior to independence, made no provision for blacks who chose an urban life, because towns were designated as white areas. As a result, though urban blacks now outnumber whites by more than four to one, blacks mostly live in rented homes in townships located some miles from city centres. The cities of Harare and Bulawayo therefore constitute studies in contrast, with impressive office buildings and quiet white suburbs partially ringed by crowded black townships. The Land Tenure Act, a more rigidly segregationist law that superseded the Land Apportionment Act in 1969, was amended in 1977, while the civil war was still being fought, to allow blacks to purchase white farms and urban property, and after the end of hostilities residential segregation began to be significantly breached.

The people. *Ethnic and linguistic composition.* More than two-thirds of the population of Zimbabwe speak Shona as their first language, while about one out of five speak Ndebele. Both Shona and Ndebele are Bantu languages; from the time of their great southward migration, Bantu-speaking tribes have populated what is now Zimbabwe for more than 10 centuries. Those who speak Ndebele are concentrated in a circle around Bulawayo, with Shona-speaking peoples beyond them on all sides—the Kalanga to the southwest, the Karanga to the east around Nyanda (formerly Fort Victoria), the Zezuru to the northeast, and the Rozwi and Tonga to the north. Generations of intermarriage have to a degree blurred the linguistic division between the Shona and Ndebele peoples.

Among the whites in Zimbabwe at independence were the descendants of the country's first European immigrants. Only about one-quarter of the adult white population, however, were born in Zimbabwe. After World War II the white population grew severalfold because of heavy immigration, and some two-thirds of the present white population have their origins in Europe, the great majority from Britain. The rest have come largely from South Africa. Of the whites living in rural areas, about one-quarter are Afrikaners.

There are several thousand Asians, forming a community that is predominantly concerned with trade. There are also Zimbabweans of mixed race, called Coloureds, who are mainly skilled and semiskilled workers.

English is the official language of government; teaching in schools is also conducted in English, except for the instruction of the youngest children in black schools.

Religion. The great majority of the black population adheres to traditional religion based on reverence for ancestors. The Shona have preserved their ancient reputation for prophecy, divination, and rainmaking; they believe in Mwari, a supreme being. The stone ruins of Great Zimbabwe are regarded as a shrine of deep religious significance, as also are parts of the Matopo Hills. In the last 50 years Christian mission schools have exercised much influence in the country, and most of the members of the first Cabinet of independent Zimbabwe were graduates of these schools. The Roman Catholic, Anglican, Methodist, Presbyterian, Baptist, and Dutch Reformed churches are represented. Because the Roman Catholic church supported nationalist aspirations, it held a position of influence in the postindependence period.

Traditional religion

Immigration and emigration. Migration has been the most important factor influencing the size and composition of the white population. Net migration figures have fluctuated in reaction to political events. In the years immediately preceding the breakup of the Federation of Rhodesia and Nyasaland, there was a net emigration of 13,000 whites; this was followed during the first 10 years of UDI by a net immigration of 40,000 (with 111,270 immigrants and 71,330 emigrants). As warfare spread after 1976, the pendulum swung again from a peak white population of 260,000 to fewer than 200,000 after independence. These net figures obscure, however, the gross turnover during 1965–79 of 132,560 immigrants and 133,864 emigrants. Even when allowance is made for the subsequent return of some emigrants, it is probable that at least half of the country's adult whites were newcomers after 1965.

Distribution of population. About one-fourth of the total population live in urban centres, nearly two-thirds of them in either Harare or Bulawayo. Among urban blacks, there is a disproportionately large number of males of working age, leaving an excess of older people, women, and children in rural areas. At least half of the black households are partly or wholly dependent on incomes earned in the wage economy.

The economy. The government of independent Zimbabwe moved cautiously to alter the pattern of management that it inherited from the white minority regime.

Private enterprise and socialism

The first budget of July 1980 was described by the finance minister as "conservative [with] a mild and pragmatic application of socialism." But the whites had passed on government machinery that included many levers of economic power. While the whites by inclination were wedded to a system of private enterprise, they had evolved a system of government intervention to support infant industries and maintain agricultural prices through marketing boards. The need to cushion the blows dealt by economic sanctions during UDI brought acceptance of the imposition of exchange and import controls.

The government raises nearly half of its revenue from personal and corporate income taxes that since 1966 have been collected on a pay-as-you-earn system. About two-fifths of government revenue comes from customs and excise duties and sales taxes, a small portion from investments, and much of the rest from government borrowing and, since independence, international aid.

The independent Zimbabwe government removed sales taxes on the staple items of food and fuel for the poorest people and extended sales taxes to travel, hotel accommodations, taxis, telecommunications, and other services. It continued the former rates of personal income tax.

The evolution of the trade union movement was some two years behind the pattern of political change by 1980. The Robert Mugabe government dealt with immediate labour problems, such as strikes for a higher minimum wage, rather than institute a thorough revision of the basic Industrial Conciliation Act of 1959. The government seemed to favour the strengthening, by mergers or amalgamation, of small unions in the same industry; the strengthening of the whole movement by the formation of a single trade-union congress from the five or six existing confederations of unions; and an arm's-length relationship of government with such a congress. Despite the large number of unions in existence, the largest sections of the labour force—the agricultural workers and domestic servants—remained outside the system.

Consolidation of trade unions

Employers' groups, such as the Associated Chambers of Commerce of Zimbabwe and the Association of Rhodesian Industries, remained influential.

Agriculture. Agriculture is the most important productive sector of the country's economy. It regularly generates about 15 percent of the gross domestic product (GDP) and some 40 percent of foreign-exchange earnings. More than one-half of the total labour force and one-fourth of the formally employed are engaged directly in agricultural activities.

The sector is divided into large-scale commercial farming, which occupies some 40 percent of the total land area and is dominated by white farmers, and small-scale farming, which is both commercial and subsistence in nature. Occupying about the same total area as the large-scale commercial sector—but on land that is considerably less fertile—smallholders have steadily increased their share of the country's total agricultural output since independence, from about one-tenth in the early 1980s to about half of total production in the early 1990s. To accelerate this trend and redress the issue of land distribution, the government has purchased many large farms and established resettlement areas on them for African farmers.

Crop production is well diversified. The most important food crop is corn, which is grown throughout Zimbabwe but does best in the well-watered northeast. Enough corn is usually produced so that Zimbabwe is able to meet its domestic demand and also export a sizable quantity.

Wheat production has grown steadily since the 1960s, but the country meets only about two-thirds of its high domestic demand and must import the rest. Other food crops include millet, sorghum, barley, cassava, peanuts (groundnuts), soybeans, bananas, and oranges.

Tobacco production

Tobacco, of which Zimbabwe is the largest producer in Africa, is the principal cash crop. It contributes about two-thirds of all agricultural export revenue and employs more than one-tenth of the work force. Three types of tobacco have traditionally been grown in the country: Virginia flue-cured, on the large commercial farms; burley, mostly by smallholders; and Turkish, of more limited extent.

Tobacco farming near Marondera, eastern Zimbabwe.

Cotton is a chief export crop and also the foundation of a large domestic textile industry. Grown by both smallholders and large commercial farmers, cotton has increased steadily in output since the UDI period, when commercial farmers were forced to diversify their production away from an overreliance on tobacco.

Sugar is grown in the southern Lowveld. It is a major export, as well as the basis for an important fuel industry, which mixes the sugar by-product ethanol with gasoline to help decrease the country's reliance on expensive imported fuels. Coffee has increased in production many times over since the early 1970s. Grown mainly in the eastern highlands between Vumba and Mount Silinda, Zimbabwe's coffees are premium mild arabicas that command a favourable price on the world market.

Cattle are the preferred livestock of the country's farmers. Beef and dairy products, produced mainly by the commercial sector, account for about one-fourth of agricultural output in most years. Since independence there has been a growing domestic demand for beef, and, as one of the few African countries allowed to export beef to the European Community, Zimbabwe has developed a significant export trade in beef as well. Sheep, goats, and pigs are raised in some areas, but their importance is minor compared to cattle. Poultry is kept largely for home use.

Industry. Although mining accounts for less than 10 percent of the GDP and provides work for only about 5 percent of the employed labour force, its significance in the economy is considerable as a major earner of foreign exchange. Direct mineral exports account for about one-third of total export earnings.

It was the prospect of great mineral wealth—comparable to the gold deposits of the Witwatersrand in neighbouring South Africa—that attracted the first permanent European settlers in the 1890s. These great expectations faded for many years after the peak of gold production was reached in 1915. By the 1950s, however, production of the chromium mines along the Great Dyke was significant, as was that of asbestos and copper. During the UDI period, the value of mining output increased. The rise in gold prices in the 1970s revived gold as the country's leading export and led to the reopening in 1979–80 of more than 100 dormant mines. Nickel mining along the Great Dyke began on a commercial scale in the late 1960s. Zimbabwe's

huge coal reserves are estimated to be about 30 billion tons, much of it desirable low-sulfur bituminous coal. Production from the major coalfields near Hwange is limited, however, by the country's capacity to transport the coal by rail, an economic necessity because of coal's bulkiness.

Manufacturing industries

Manufacturing generates about one-fourth of the country's GDP, a proportion much greater than in any other black African state. From 1954 to 1963, then Southern Rhodesia was able to rely on the resources and larger market of the Federation of Rhodesia and Nyasaland for a 150 percent increase in manufacturing output. Then, after UDI in 1965, hundreds of new manufacturing projects were begun in an effort to defeat economic sanctions by import substitution. Because of the diversity in manufacturing that developed then, Zimbabwe today is able to provide nearly 90 percent of the manufactured goods used in the country. The export market for Zimbabwe's manufacturers, however, is hindered by a lack of foreign exchange in neighbouring countries.

Coal is the nation's primary energy source. A growing percentage of the coal utilized is transformed first into electricity, however, as new thermal generating plants fueled by coal are constructed. Electricity and water production contributes about 3 percent of the GDP. Its principal users are industries, mines, and farms. Electrification of the railways was begun in 1980 (coal and diesel remain the major energy sources for rail transport, however), and there has also been considerable electrification of low-cost housing in urban townships. Less than half the black homes in Bulawayo and Harare had their own electricity at the time of independence.

Apart from South Africa, Zimbabwe generates more electricity per capita than any country in sub-Saharan Africa. About one-third of the electric power is generated at the huge Kariba Dam on the Zambezi River. Most of the remainder is produced thermally.

Finance and trade. Finance and insurance services contribute about 6 percent of the GDP. The Reserve Bank of Zimbabwe, located in Harare, is the country's central bank. It is the sole bank of issue and administers all monetary and exchange controls. There are also private and government-sponsored commercial banks, a development finance bank, several merchant banks and discount houses, and a post office savings bank for individual savings accounts. The Zimbabwe Stock Exchange deals in both government securities and the securities of privately owned companies.

Economic sanctions during UDI, which had been imposed by stages from 1966 to 1968 on both imports and exports, were lifted in December 1979. They had been widely breached, particularly in mineral exports and in the supply of petroleum, but they nevertheless strongly affected certain commodities, such as tobacco exports. Although the trade surplus was diminished in 1979 by the rise in oil prices, the value of exports still outpaced that of imports. In the 1980s, Zimbabwe showed slow but steady growth in its trade surplus, as its unusually high level of export diversity proved able to weather changes in world demand for its commodities.

Trade surplus

Tobacco is the major export, accounting for about one-fifth of all export earnings, followed by gold, metal alloys, cotton, textiles, asbestos, corn, and sugar. The principal imports are fuels and petroleum products, aircraft and spare parts, chemicals, iron and steel plates, and miscellaneous manufactured goods. Zimbabwe's major trading partners are the United Kingdom, South Africa, Germany, Japan, and the United States.

Transportation. The main road system, which is excellent, generally follows the line of white settlement along the spine of the country, with two branches north to Victoria Falls and Kariba and a network fanning out from Nyanda, close to the Great Zimbabwe ruins. Wartime operations brought an improvement in certain areas, including the construction of strategic roads in the eastern highlands and near the Zambian border. The roads in white farming areas and the gravel and earth roads in the Tribal Trust Lands received barely adequate maintenance, however.

Zimbabwe has one of the densest rail networks in sub-Saharan Africa. The railway closely follows the main road network; its single track has a gauge of three feet six inches. The country has rail links with South Africa to the south and Zambia to the north. Two lines connect with lines through Mozambique to give landlocked Zimbabwe access to the ports of Maputo and Beira.

Air Zimbabwe replaced Air Rhodesia, a government-backed company that had operated only within Rhodesia and to and from South Africa. The international airport at Harare has one of the longest civil runways in the world. Seven other airports—at Bulawayo, Kariba, Gweru, Masvingo, Hwange, Buffalo Range, and Victoria Falls—can accommodate medium-size jet aircraft.

Administration and social conditions. *Government.* The constitution of Zimbabwe, which was written in London during September–December 1979 and which took effect at independence on April 18, 1980, secured majority rule for Zimbabweans. Under the constitution, white voters, registered on a separate roll, elected 20 of the 100 members of the House of Assembly. Although these members no longer can veto constitutional amendments, a unanimous vote was required during the first 10 years to alter the Declaration of Rights, which stipulates (among other matters) that, if land is acquired for settlement schemes, there must be "prompt payment of adequate compensation . . . remittable within a reasonable time to any country outside Zimbabwe." The British insisted that there be a constitutional head of state, a president elected by the House of Assembly, and an executive prime minister and that citizenship of Zimbabwe be automatically available to anyone who was (or had the qualifications to be) a citizen of Rhodesia immediately before independence. The former Senate of 40 members was abolished with a constitutional amendment in 1990, and 50 members were added to the House of Assembly. One hundred twenty seats in the assembly were elective, 10 secured for traditional chiefs, and 8 for the provincial governors, with the remaining 12 to be appointed by the president.

The 1980 constitution

At the time of independence, whites controlled the municipal councils, but legislation was soon introduced to amalgamate each municipal council with the council of its surrounding township, and, for the first time, black mayors were elected in 1981. Local government elections in rural areas replaced the old apparatus of district commissioners with a party-based council structure.

Justice. Under the constitution, the four-member Judicial Service Commission advises the president on the appointment of judges to the High Court. High Court judges may not be removed from office except for misconduct or incapacity. The first black lawyer was appointed a High Court judge in 1980. In addition to magistrates who preside over criminal and civil litigation, other courts adjudicate on matters of African law and custom.

Education. The dismantling of Rhodesia's segregated system of schooling began less than two years before independence. The minority government had concentrated upon providing compulsory (and virtually free) education for white children between the ages of 5 and 15 and had left the schooling of black children in the hands of missionaries. In 1950 there were only 12 government schools for blacks, compared with 2,230 mission and independent schools.

After independence, priority was given to upgrading the nation's school system. Many new schools were built in the drive toward free primary education for all. In the decade following independence, Zimbabwe achieved one of the highest primary school enrollment rates in Africa, with more than 90 percent of all children of primary school age attending school. Primary education begins at age seven, lasts for seven years, and has been compulsory since 1987. At least one rural secondary school has been established in each of the country's 55 districts. The University of Zimbabwe, founded in 1955 at Harare, is the country's primary institution of higher education.

Primary school enrollment

Health and welfare. Before 1980, health services were biased toward curative medicine in central hospitals. Missionaries had the major responsibility for running rural clinics and small hospitals. After independence, health allocations were increased.

As in other Third World countries, the burden of disease

is heaviest on Zimbabwe's youngest children. The infant mortality rate in malarial parts of the Zambezi valley has been as high as 300 per 1,000, and the rate is thought to lie between 120 and 220 per 1,000 as a whole. Measles, malaria, diarrheal diseases, and pneumonia are major causes of death. AIDS, however, has become the major health threat to Zimbabweans in the 21st century, with fully one-fourth of the adult population being infected. Improved nutrition is increasingly seen as the most important health need.

Cultural life. The most famous of Rhodesian-bred writers, Doris Lessing, settled in England in 1949. In some contrast, the nationalist struggle prompted a renaissance of Shona culture. A forerunner of this renaissance (and a victim of the liberation struggle) was Herbert Chitepo, both as abstract painter and epic poet. Stanlake Samkange's novels reconstruct the Shona and Ndebele world of the 1890s, while those of the much younger Charles Mungoshi explore the clash of Shona and Western cultures in both the Shona and English languages. Folk traditions have survived in dance and pottery. The revival of sculpture has drawn on tribal religion and totems to produce some remarkable works, particularly those of Takawira and the Tengenenge school of craftsmen who sculpt in hard serpentine. (C.W.S./Ed.)

Cultural renaissance

For statistical data on the land and people of Zimbabwe, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The remains of Stone Age cultures dating to 500,000 years ago have been found in Zimbabwe, and it is thought that the San, who still survive mostly in the Kalahari desert of Botswana, are the last descendants of these original inhabitants of southern and central Africa. They were driven into the desert by Bantu-speaking peoples during the long migrations from the north in the course of which the Bantu-speaking peoples populated much of Africa from Lake Chad to present-day South Africa. The first Bantu are thought to have reached Zimbabwe between the 5th and 10th centuries AD. The stone ruins of Zimbabwe date mainly from about the 9th century, although the most elaborate belong to a period after the 15th century and are of Bantu origin.

Portuguese exploration. The Portuguese, who arrived on the east coast of Africa at the end of the 15th century, dreamed of opening up the interior and establishing a route to connect their eastern settlements with Angola in the west. The first European to enter Zimbabwe was probably Antônio Fernandes, who tried to cross the continent and reached the neighbourhood of Que Que (now Kwekwe). Nearly 50 years later the "emperor" Mwene Matapa was baptized by a Jesuit father, and in 1569 an abortive Portuguese military expedition entered the interior in search of gold.

A second great movement of the Bantu peoples began in 1830, this time from the south. To escape from the power of the great Zulu chief Shaka, three important groups fled northward, one of them the Ndebele, who carved out a kingdom. The Ndebele were warriors and pastoralists, in the Zulu tradition, and under their formidable chief Mzilikazi they mastered and dispossessed the weaker tribes, known collectively as Shona (Mashona), who were sedentary, peaceful tillers of the land. For more than half a century, until the coming of European rule, the Ndebele continued to enslave and plunder the Shona. During this period, however, British and Afrikaner hunters, traders, and prospectors had begun to move up from the south, and with them came the missionaries. Robert Moffat visited Mzilikazi in 1857, and this meeting led to the establishment in 1861 of the first mission to the Ndebele by the London Missionary Society.

The British South Africa Company. In South Africa Cecil Rhodes formed the British South Africa Company, which received its charter in October 1889. Its objects were (1) to extend the railway from Kimberley northward to the Zambezi, (2) to encourage immigration and colonization, (3) to promote trade and commerce, and (4) to secure all mineral rights, in return for guarantees of protection and security of rights to the tribal chiefs.

In 1890 a pioneer column set out from Bechuanaland and reached the site of the future capital of Rhodesia without incident on September 12. There the new arrivals settled and began to lay claim to prospecting rights. The Ndebele resented this European invasion, and in 1893 they took up arms, being defeated only after months of strenuous fighting. Lobengula, Mzilikazi's son and successor, fled, and the company assumed administrative control of Matabeleland. In 1895 many of the pioneers were persuaded to take part in the Jameson Raid into the Transvaal and were captured and sent to England for trial. In the same year, the company-administered territories, which had previously been loosely known as Zambesia, were formally named Rhodesia by proclamation. In 1896 the Ndebele rose again. Returning from London, Rhodes met with the Ndebele chiefs and persuaded them to make peace. The Shona had at first accepted the Europeans, but they too became rebellious, and the whole country was not pacified until 1897.

European settlement

Economic and political development. By 1892 about 1,500 settlers from the south had arrived in Rhodesia. The railway reached Bulawayo in 1896 and Victoria Falls in 1904. By the following year there were 12,500 settlers in the country, and in 1909 gold exports were worth more than £2,500,000. Agricultural development, however, was slower, and it was not until 1907 that steps were taken to facilitate the acquisition of land. By 1911 nearly £35,000 worth of tobacco was being exported annually, and the European population had risen to 23,600.

From the earliest years, the settlers had demanded representation on the Legislative Council, which in 1903 comprised seven company officials and seven elected representatives of the settlers. In 1907 the settlers were given a majority of seats. In 1914, when the 25-year term of the company's charter was due to expire, the settlers, faced with the alternative of joining the Union of South Africa, asked for the continuation of the charter pending the grant of self-government. The British government therefore extended the charter for 10 years, with the proviso that self-government could be granted earlier if the settlers showed themselves capable of administering the country unaided.

Self-government. Immediately after World War I the pressure for self-government was resumed, and a royal commission was appointed to consider the future of the territory. As a result of the commission's report, a referendum of the electors among the 34,000 Europeans in the country was held in 1922; the choice was between entry into the Union of South Africa as its fifth province and full internal self-government. In spite of the offer of generous terms by the Union's prime minister, General Jan C. Smuts, a majority voted for self-government. On Sept. 12, 1923, Southern Rhodesia was annexed to the crown and became a self-governing colony. The British government retained control of external affairs and a final veto in respect to legislation directly affecting Africans.

The interwar period was one of material progress, with the development of a reasonably prosperous economy based on copper, gold, and other minerals, corn, tobacco, and cattle. By 1953 Southern Rhodesia had a European population of 157,000 and an annual revenue of more than £28 million.

The policy of Sir Godfrey Huggins (later Lord Malvern), who served as prime minister of Southern Rhodesia for 20 years, was to build a society in accord with Rhodes's dictum of "equal rights for all civilized men," one in which merit and not colour should be the test of political and economic advancement. He believed that political power should not be given to the Africans until they were sufficiently experienced to know how to exercise it in cooperation with the Europeans and thus to maintain the economic development built up over the years.

Lord Malvern's governing principles

A second principle in which Lord Malvern and most other Europeans in Southern Rhodesia and Northern Rhodesia (later Zambia) profoundly believed was that the two countries should be joined together, both for their mutual economic benefit and to ensure the establish-

ment of a powerful state based on British culture and traditions. Malvern failed to secure their amalgamation, but he supported the federation of Southern Rhodesia, Northern Rhodesia, and Nyasaland (later Malawi) when that solution was eventually accepted by the British in 1953.

Federation. In 1957 a new electoral law was passed providing for a common roll of voters (the "A" roll, composed only of whites) with a special roll for those with lower qualifications (the "B" roll, a tiny minority of the blacks). At the same time, there was growing political consciousness among the African population, together with increasing hostility to the idea of federation. Joshua Nkomo was one of the fiercest opponents of federation as the local leader of the African National Congress, and when that organization was banned he became president of the National Democratic Party in 1960. It, too, was soon banned, and he formed the Zimbabwe African People's Union (ZAPU), which in turn was banned in 1962. In 1963 Robert Mugabe broke with ZAPU to join the Zimbabwe African National Union (ZANU) and thereby split African support along ethnic lines—Nkomo retained the Ndebele ethnic minority (mostly in the Matabeleland region), while Mugabe garnered the Shona ethnic majority.

Formation of ZAPU and ZANU

In June 1962 the UN General Assembly called for a more liberal constitution for the territory. The election of December 1962, during which the 1961 constitution came into force, was boycotted by the African nationalists. The ruling United Federal Party was defeated by the more conservative Rhodesian Front (RF), and Winston Field became prime minister. At the end of 1963 the federation was dissolved, and Southern Rhodesia reverted to its former status as colony. (K.Br./K.In.)

Rhodesia and the UDI. The goal of the RF was Rhodesian independence under guaranteed minority rule. Field was replaced as prime minister in April 1964 by his deputy, Ian Smith. The RF swept all A-roll seats in the 1965 election, and Smith used this parliamentary strength to tighten controls on the political opposition. After several attempts to persuade Britain to grant independence, Smith's government announced the Unilateral Declaration of Independence (UDI) on Nov. 11, 1965.

British response to the UDI

Britain declined to respond to the UDI with force, instead attempting economic tactics such as ending the link between sterling and the Rhodesian currency and seizing assets. Smith's government countered by defaulting on its (British-guaranteed) debts, leaving the British liable while at the same time balancing its budget. The United Nations imposed economic sanctions in 1968, but these were only partly successful; some strategic minerals, especially chromium, were exported to willing buyers in Europe and North America, further strengthening the economy.

Unsuccessful negotiations with Britain continued. A 1971 proposal to lessen restrictions on the opposition led to the creation of a third nationalist movement, the United African National Council (UANC), led by the Methodist bishop Abel Muzorewa. Unlike ZAPU and ZANU—both banned and operating only from exile in Zambia and Mozambique, respectively—UANC was able to organize inside Rhodesia and held talks with the government during the 1970s. During the early 1970s ZAPU and ZANU had sporadically organized raids into Rhodesia, but in December 1972 the violence of the conflict intensified after a ZANU attack in the northeast. The Zambia-Rhodesia border was closed in 1973, but Mozambican independence in 1975 provided a valuable base of operations for ZANU, which had close links to the Frelimo government.

The white Rhodesian government was thus under diplomatic, military, and, increasingly, economic pressure for a settlement. The 1976 rapprochement between Nkomo and Mugabe led to the formation of the Patriotic Front (PF), which received "frontline" support from Rhodesia's majority-ruled neighbours. The fighting escalated in both area and intensity, and the emergency measures adopted by the government to counter it also served to increase antigovernment feeling. By 1979 the combination of pressures had forced Smith to accept the necessity of an "internal settlement."

Independence. A 1978 agreement with internal black leaders, including Muzorewa, promised elections for a transitional government that would provide for both black enfranchisement and protection of white political and economic interests. The UANC won a clear majority of the seats allotted to blacks in the April 1979 election, and the country adopted the name Zimbabwe. Without PF participation or support for Muzorewa's new government, however, Zimbabwe was unable to end the warfare. Diplomatic recognition of the new government was not forthcoming given the stalemate; after talks between Muzorewa, Mugabe, and Nkomo in London in late 1979, Britain briefly retook control of Southern Rhodesia as a colony until a new round of elections was held in February 1980. Of the 80 contested black seats, ZANU won 57, ZAPU 20, and the UANC 3. Mugabe became the first prime minister as Zimbabwe achieved an internationally recognized independence on April 18, 1980.

Mugabe's government

Mugabe's new government moved deliberately to redress inequalities of race and class, redistribute land, and promote economic development, with a one-party socialist state as its long-term goal. During the 1980s, drought and white emigration badly damaged the economy, which was already strained by the need for massive government spending in the long-neglected areas of black education, health, and social services. In 1982 Mugabe charged that Nkomo was plotting a coup and forced him out of office, while arresting other leaders of ZAPU. Nkomo's supporters in the Matabeleland region retaliated, precipitating a civil war. Fighting did not cease until Mugabe and Nkomo reached an agreement in December 1987 whereby ZAPU was subsumed into ZANU-PF, Mugabe became the country's first executive president, and Nkomo became one of the nation's two vice presidents. Mugabe was reelected in 1990, 1996, and 2002.

From the beginning of independence in 1980, the government struggled with the issue of land reform. Some 4,000 white farmers controlled about one-third of Zimbabwe's arable land, and squatters, incited by government promises and the lack of police protection, partially took over hundreds of white-owned farms. Nevertheless, public support for the farmers and opposition to Mugabe's increasingly autocratic rule were evidenced in February 2000 by the defeat of a referendum for a new constitution that would have given him the power to confiscate white-owned farms without compensation. Land seizure accelerated in mid-2002, however, as Mugabe was reelected yet again, albeit in controversial elections, and laws were enacted that made it easier to seize land. Mugabe's land-reform policies and his 1998 intervention in the civil war in Congo (Kinshasa), purportedly to protect his personal investments, provoked international outrage and the suspension of some economic aid for Zimbabwe. High unemployment and high rates of inflation also created problems.

In addition to severe political and economic turmoil, the spread of AIDS in Zimbabwe reached epidemic proportions. By the beginning of the 21st century, one in four adult Zimbabweans was infected, life expectancy had fallen to below 40 years, and hundred of thousands of children had been left orphans. (Ed.)

For later developments in the history of Zimbabwe, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

General works. *Africa South of the Sahara* (annual) and *Africa Contemporary Record* (annual) contain essays on southern African countries. The region may be seen in its larger context in ROLAND OLIVER and MICHAEL CROWDER (eds.), *The Cambridge Encyclopedia of Africa* (1981); while the *Standard Encyclopaedia of Southern Africa,* 12 vol. (1970–76), focuses on the region.

Physical and human geography. *The land:* ALAN B. MOUNTJOY and DAVID HILLING, *Africa: Geography and Development* (1988); and A.T. GROVE, *The Changing Geography of Africa* (1989), include discussions of regional geography. N. LANCASTER, *The Namib Sand Sea: Dune Forms, Processes, and Sediments* (1989), reviews the geomorphology of the Namib; while DAVID S.G. THOMAS and PAUL A. SHAW, *The Kalahari Environment* (1991), brings together work on the Kalahari desert. P.D. TYSON, *Climatic Change and Variability in Southern Africa*

(1986), discusses weather patterns. M.J.A. WERGER and A.C. VAN BRUGGEN (eds.), *Biogeography and Ecology of Southern Africa,* 2 vol. (1978), provides comprehensive coverage of the plant and animal life of the region. ANDREW MILLINGTON *et al., Biomass Assessment: Woody Biomass in the SADCC Region* (1989), assesses land cover and woody biomass resources in southern Africa for fuelwood energy planning. BARRY MUNSLOW *et al., The Fuelwood Trap: A Study of the SADCC Region* (1988), analyzes fuelwood collection, land degradation, and its socio-economic effects over much of southern Africa. (A.C.Mi.)

The people: Studies of southern African peoples are contained in GEORGE MURDOCK, *Africa: Its Peoples and Their Culture History* (1959); HAROLD K. SCHNEIDER, *The Africans: An Ethnological Account* (1981); JOCELYN MURRAY (ed.), *Cultural Atlas of Africa* (1981); CARMEL SCHRIRE (ed.), *Past and Present in Hunter-Gatherer Studies* (1984); JAN VANSINA, *Paths in the Rainforests: Toward a History of Political Tradition in Equatorial Africa* (1990), and "Western Bantu Expansion," *Journal of African History,* 25(2):129–145 (1984); EDWIN N. WILMSEN, *Land Filled with Flies: A Political Economy of the Kalahari* (1989), a challenge to stereotypes of hunter-gatherers in the Kalahari; RICHARD B. LEE and IRVEN DEVORE (eds.), *Kalahari Hunter-Gatherers: Studies of the !Kung San and Their Neighbors* (1976); ANTHONY TRAILL, "The Languages of the Bushmen," in PHILLIP V. TOBIAS (ed.), *The Bushmen: San Hunters and Herders of Southern Africa* (1978), pp. 137–147; and LEROY VAIL (ed.), *The Creation of Tribalism in Southern Africa* (1989). REGINALD H. GREEN *et al., Children on the Front Line: The Impact of Apartheid, Destabilization, and Warfare on Children in Southern and South Africa,* 3rd ed. (1989), reviews the position of children in the region in the context of war with reference to lives lost and GDP eroded as well as health service and nutrition issues. (J.R.De.)

The economy: Aspects of the southern African economy, including agriculture, policies, and development, are treated in *Survey of Economic and Social Conditions in Africa* (annual), a United Nations publication; RALPH A. AUSTEN, *African Economic History: Internal Development and External Dependency* (1987); ROGER RIDDELL *et al., Manufacturing Africa: Performance & Prospects of Seven Countries in Sub-Saharan Africa* (1990); HANS BINSWANGER and PRABHU PINGALI, "Technological Priorities for Farming in Sub-Saharan Africa," *Journal of International Development,* 1(1):46–65 (January 1989); *Sub-Saharan Africa: From Crisis to Sustainable Growth: A Long-term Perspective Study* (1989), a World Bank publication; ALLAN LOW, *Agricultural Development in Southern Africa: Farm-Household Economics and the Food Crisis* (1986); and JOSEPH HANLON, *SADCC in the 1990s: Development on the Front Line* (1989). ROBIN PALMER and NEIL PARSONS (eds.), *The Roots of Rural Poverty in Central and Southern Africa* (1977), contains key essays on the region's political economy in the 19th and 20th centuries. (J.B.S./H.Za.)

History. *General works:* J.D. FAGE and ROLAND OLIVER, *The Cambridge History of Africa,* 8 vol. (1975–86), has long chapters on southern Africa by leading authorities and places southern Africa in the context of African history; a more condensed attempt is provided by PHILIP CURTIN *et al., African History* (1978). UNESCO INTERNATIONAL SCIENTIFIC COMMITTEE FOR THE DRAFTING OF A GENERAL HISTORY OF AFRICA, *General History of Africa,* 8 vol. (1981–93), is an international collaborative effort that locates African people and their experience at the centre and portrays African contact with Middle Eastern, Asian, and Euro-American people in the larger context of African history. NEIL PARSONS, *A New History of Southern Africa,* 2nd ed. (1993), is an introductory text dealing with the region as a whole. DAVID BIRMINGHAM and PHYLLIS MARTIN (eds.), *History of Central Africa,* 2 vol. (1983), contains essays on topics in central and southern African history. A.J. WILLS, *An Introduction to the History of Central Africa: Zambia, Malawi, and Zimbabwe,* 4th ed. (1985), is still a useful guide to British south-central Africa. LEROY VAIL and LANDEG WHITE, *Power and the Praise Poem: Southern African Voices in History* (1991), signals a major turning point in the interpretation of southern Africa's past by insisting on the centrality of African interpretations through poetry, performance, and other oral expressions. Current research can be found in such specialist journals as *Journal of African History* (3/yr); *Journal of Southern African Studies* (quarterly); and *African Affairs* (quarterly).

Southern Africa to 1800: D.W. PHILLIPSON, *African Archaeology,* 2nd ed. (1993), is an introductory text with coverage ranging from prehistoric times to European contact. Overviews of early societies are found in R.R. INSKEEP, *The Peopling of Southern Africa* (1978); and D.W. PHILLIPSON, *The Later Prehistory of Eastern and Southern Africa* (1977), which both deal with the Late Stone Age and the Iron Age; and PETER S. GARLAKE, *The Kingdoms of Africa* (1978, reissued 1990), an excellent introduction to the Iron Age in southern Africa, and *Great Zimbabwe* (1973). GRAHAM CONNAH, *African Civilizations: Precolonial Cities and States in Tropical Africa: An Archaeological Perspective* (1987), discusses the period of Great Zimbabwe's influence (*c.* 1250–1450) and Bantu occupation of east coast areas from the end of the 1st millennium AD. MARTIN HALL, *The Changing Past: Farmers, Kings, and Traders in Southern Africa, 200–1860* (1987; also published as *Farmers, Kings, and Traders,* 1990), offers an overview of southern African history and archaeology, starting with the first settlement of the subcontinent. Portuguese ventures in west-central and east-central Africa are chronicled in EDWARD A. ALPERS, *Ivory and Slaves* (also published as *Ivory & Slaves in East Central Africa,* 1975); DAVID BIRMINGHAM, *Trade and Conflict in Angola: The Mbundu and Their Neighbours Under the Influence of the Portuguese, 1483–1790* (1966); JOSEPH C. MILLER, *Kings and Kinsmen: Early Mbundu States in Angola* (1976); C.R. BOXER, *Race Relations in the Portuguese Colonial Empire, 1415–1825* (1963, reprinted 1985); JAN VANSINA, *Kingdoms of the Savanna* (1966); and PHYLLIS M. MARTIN, *The External Trade of the Loango Coast, 1576–1870* (1972).

Southern Africa, 1800–c. 1900: DONALD DENOON and BALAM NYEKO, *Southern Africa Since 1800,* new ed. (1984), is a good overview. The Mfecane and its effects are analyzed in J.D. OMER-COOPER, *The Zulu Aftermath* (1966, reissued 1978), the standard account covering southern, central, and eastern Africa, the findings of which have been modified by detailed regional research. A stimulating account of the colonial period can be found in MARTIN CHANOCK, *Law, Custom, and Social Order: The Colonial Experience in Malawi and Zambia* (1985). British policy and the expansion of European settlement south of the Limpopo River is addressed in JOHN S. GALBRAITH, *Reluctant Empire: British Policy on the South African Frontier, 1834–1854* (1963, reprinted 1978); and DAVID WELSH, *The Roots of Segregation: Native Policy in Colonial Natal, 1845–1910* (1971). Attempts to look at the African side in this period include ROBERT ROSS, *Adam Kok's Griquas: A Study in the Development of Stratification in South Africa* (1976); and NORMAN ETHERINGTON, *Preachers, Peasants, and Politics in Southeast Africa, 1835–80: African Christian Communities in Natal, Pondoland, and Zululand* (1978). The era of mineral discoveries and the scramble for southern Africa are the subject of C.W. DE KIEWIET, *The Imperial Factor in South Africa* (1937, reissued 1966); applicable chapters in RONALD ROBINSON, JOHN GALLAGHER, and ALICE DENNY, *Africa and the Victorians: The Official Mind of Imperialism,* 2nd ed. (1981); D.M. SCHREUDER, *The Scramble for Southern Africa, 1877–1895* (1980), a synthesis of the subject; J.S. MARAIS, *The Fall of Kruger's Republic* (1961); and DONALD DENOON, *A Grand Illusion: The Failure of Imperial Policy in the Transvaal Colony During the Period of Reconstruction, 1900–1905* (1973). A critique of the literature on British policy in South Africa can be found in ANTHONY ATMORE and SHULA MARKS, "The Imperial Factor in South Africa: Towards a Reassessment," *Journal of Imperial and Commonwealth History* (October 1974). The scramble in central Africa and the establishment of colonial society are dealt with in JOHN S. GALBRAITH, *Crown and Charter: The Early Years of the British South Africa Company* (1974); and IAN PHIMISTER, "Rhodes, Rhodesia, and the Rand," *Journal of Southern African Studies,* vol. 1, no. 1 (1974), a radical reinterpretation of the connections. Events in Angola and Mozambique are outlined in MALYN NEWITT, *Portugal in Africa: The Last Hundred Years* (1981), a lucid overview with a synopsis of the earlier period of Portuguese rule; and GERVASE CLARENCE-SMITH, *The Third Portuguese Empire, 1825–1975* (1985), an overview from an economic perspective. German activity in South West Africa is discussed in PROSSER GIFFORD, W.M. ROGER LOUIS, and ALISON SMITH (eds.), *Britain and Germany in Africa* (1967); and HORST DRECHSLER, *Let Us Die Fighting: The Struggle of the Herero and Nama Against German Imperialism (1884–1915)* (1980; originally published in German, 1966).

Southern Africa, c. 1900 to the present: MARTIN CHANOCK, *Britain, Rhodesia, and South Africa, 1900–45* (also published as *Unconsummated Union,* 1977), contains a masterly account of interregional politics. Coverage of more recent events from differing viewpoints can be found in BASIL DAVIDSON, JOE SLOVO, and ANTHONY WILKINSON, *Southern Africa: The New Politics of Revolution* (1976), on the struggle in the Portuguese colonies, South Africa, and Rhodesia following the Portuguese coup of 1974. Coverage of various countries during this period is found in LORD HAILEY (WILLIAM MALCOLM HAILEY, BARON HAILEY), *Native Administration in the British African Territories,* 5 vol. (1950–53, reprinted 5 vol. in 3, 1979), and *The Republic of South Africa and the High Commission Territories* (1963, reprinted 1982); RICHARD GRAY, *The Two Nations: Aspects of the Development of Race Relations in the Rhodesias and Nyasaland* (1960, reprinted 1974); ROBERT I. ROTBERG, *The Rise of Nationalism in Central Africa: The Making of*

Malawi and Zambia, 1873–1964 (1965); CHARLES PERRINGS, *Black Mineworkers in Central Africa: Industrial Strategies and the Evolution of an African Proletariat in the Copperbelt, 1911–41* (1979); and PATRICK KEATLEY, *The Politics of Partnership* (1963), on Rhodesia and Nyasaland. (Sh.M.)

Angola. Overviews are provided by DOUGLAS L. WHEELER and RENÉ PÉLISSIER, *Angola* (1971); and THOMAS COLLELO (ed.), *Angola, a Country Study,* 3rd ed. (1991). MANFRED KUDER, *Angola: eine geographische, soziale, und wirtschaftliche Landeskunde* (1971), is the standard geographic text. LAWRENCE W. HENDERSON, *A Igreja em Angola* (1990), surveys the Christian groups. The best up-to-date information on politics and the economy is provided by two publications from the Economist Intelligence Unit, *Country Profile: Angola, São Tomé & Príncipe* (annual), and *Country Report: Angola, São Tomé & Príncipe* (quarterly). *Angola, an Introductory Economic Review* (1991), a World Bank publication, is also useful.

LAWRENCE W. HENDERSON, *Angola: Five Centuries of Conflict* (1979), provides a broad treatment. PHYLLIS M. MARTIN, *Historical Dictionary of Angola* (1980), includes a bibliography. DAVID BIRMINGHAM, *Central Africa to 1870: Zambezia, Zaïre, and the South Atlantic* (1981), synthesizes prehistory and the early Portuguese presence. More on the Portuguese presence can be found in GERALD J. BENDER, *Angola Under the Portuguese: The Myth and the Reality* (1978). Analyses of the slave trade include JOSEPH C. MILLER, *Way of Death: Merchant Capitalism and the Angolan Slave Trade, 1730–1830* (1988), a masterful work; and W.G. CLARENCE-SMITH, *Slaves, Peasants, and Capitalists in Southern Angola, 1840–1926* (1979). RENÉ PÉLISSIER, *Les Guerres grises: résistance et révoltes en Angola, 1845–1941* (1977), reviews modern history from a military and political perspective. JOHN A. MARCUM, *The Angolan Revolution,* 2 vol. (1969–78), analyzes the liberation movements up to independence. HENRIQUE GUERRA, *Angola: estrutura económica e classes sociais,* 4th ed. (1979), is an essay on late colonial reforms. DAVID BIRMINGHAM, *Frontline Nationalism in Angola & Mozambique* (1992), focuses on the period from 1961 to 1975. F.W. HEIMER, *The Decolonization Conflict in Angola, 1974–76: An Essay in Political Sociology* (1979), provides the best guide to decolonization. W.G. CLARENCE-SMITH, "Class Structure and Class Struggles in Angola in the 1970s," *Journal of Southern African Studies,* 7(1):109–126 (October 1980), analyzes the beginnings of the civil war. FRED BRIDGLAND, *Jonas Savimbi: A Key to Africa* (1986), views the civil war from a UNITA perspective; while KEITH SOMERVILLE, *Angola: Politics, Economics, and Society* (1986), does the same from an MPLA perspective. WILLIAM MINTER, *Apartheid's Contras: An Inquiry into the Roots of War in Angola and Mozambique* (1994), also discusses the recent civil wars. (W.G.C.-S./Ed.)

Botswana. ROBSON M.K. SILITSHENA and G. MCLEOD, *Botswana: A Physical, Social, and Economic Geography* (1989), briefly but authoritatively covers all aspects of Botswana's geography from climate to patterns of human settlement. *Botswana Notes and Records* (annual), is a scholarly journal covering research in natural and social science and in humanities. BOTSWÁNA SOCIETY, *The Botswana Society Social Studies Atlas* (1988), contains maps of the country's physical and human geography and history as well as environmental and thematic maps of the southern African region. Studies of the plant and animal life include RONALD DANIEL AUERBACH, *The Amphibians and Reptiles of Botswana* (1987); MICHAEL MAIN, *Kalahari: Life's Variety in Dune and Delta* (1987), a popular survey of recent scientific research on the natural and human aspects of the desert from prehistory to future prospects; KENNETH NEWMAN, *Newman's Birds of Botswana* (1989); MARK OWENS and DELIA OWENS, *Cry of the Kalahari* (1984); KAREN ROSS, *Okavango: Jewel of the Kalahari* (1987), an illustrated description of wetland wildlife and the environment of the delta; and DAVID S.G. THOMAS and PAUL A. SHAW, *The Kalahari Environment* (1991), an extensive scholarly study of the Kalahari Basin, detailing present geology, tectonics, climate, and vegetation and discussing the findings of recent research on prehistoric lakes and rivers from the Cretaceous to the Quaternary as well as presenting current debate on land use and water resources. TORE JANSON and JOSEPH TSONOPE, *Birth of a National Language: The History of Setswana* (1991), traces the development of a standardized national dialect beginning in the 19th century to the increasing distinction since the 1960s between Botswana and South African official Setswana. Studies of the people include LAURENS VAN DER POST and JANE TAYLOR, *Testament to the Bushmen* (1984); BESSIE HEAD, *Serowe, Village of the Rainwind* (1981), a novelist's view of local social history based on interviews; and DIANA WYLIE, *A Little God: The Twilight of Patriarchy in a Southern African Chiefdom* (1990), a historian's view of disappearing chieftainship. The economy is addressed in CHRISTOPHER COLCLOUGH and STEPHEN MCCARTHY, *The Political Economy of Botswana: A Study of Growth and Distribution* (1980), a scholarly survey of economic development, 1965–77; CHARLES HARVEY and STEPHEN R. LEWIS, JR., *Policy Choice and Development Performance in Botswana* (1990), an analysis of the successful government negotiation of terms with mining companies and the management of the resulting financial surplus; JOHN HOLM and PATRICK MOLUTSI (eds.), *Democracy in Botswana* (1989), a collection of symposium proceedings that emphasizes the adaptation of indigenous institutions, the roles of bureaucrats and foreign capital, the group rights of cultural minorities, and the efficacy of regular general elections; and LOUIS A. PICARD, *The Politics of Development in Botswana: A Model for Success?* (1987).

Historical treatments include THOMAS TLOU and ALEC CAMPBELL, *History of Botswana* (1984); FRED MORTON, ANDREW MURRAY, and JEFF RAMSAY, *Historical Dictionary of Botswana,* new ed. (1989), a reference work with an excellent bibliography; FRED MORTON and JEFF RAMSAY (eds.), *The Birth of Botswana: A History of the Bechuanaland Protectorate from 1910 to 1966* (1987), on the rise and fall of powerful local sovereignties headed by chiefs under colonial rule; MICHAEL CROWDER, *The Flogging of Phineas McIntosh: A Tale of Colonial Folly and Injustice: Bechuanaland, 1933* (1988); LOUIS A. PICARD (ed.), *The Evolution of Modern Botswana* (1985), the country's administrative history from the 1930s to postcolonial development; MICHAEL DUTFIELD, *A Marriage of Inconvenience: The Persecution of Ruth and Seretse Khama* (1990); JACK PARSON (ed.), *Succession to High Office in Botswana: Three Case Studies* (1990); and JACK PARSON, *Botswana: Liberal Democracy and the Labor Reserve in Southern Africa* (1984). (Ne.P.)

Lesotho. COLIN MURRAY, *Families Divided: The Impact of Migrant Labour in Lesotho* (1981), studies the effects of shifting labour migration on family life in three different villages. WILLIAM F. LYE and COLIN MURRAY, *Transformations on the Highveld: The Tswana & Southern Sotho* (1980), studies these people who live in Lesotho, Botswana, and central South Africa. The economy and government policies are discussed in JOHN E. BARDILL and JAMES H. COBBE, *Lesotho: Dilemmas of Dependence in Southern Africa* (1985); and JAMES FERGUSON, *The Anti-Politics Machine: "Development," Depoliticization, and Bureaucratic Power in Lesotho* (1990).

The history of the country is analyzed by STEPHEN J. GILL, *A Short History of Lesotho: From the Late Stone Age Until the 1993 Elections* (1993); ROBERT C. GERMOND (compiler and trans.), *Chronicles of Basutoland* (1967), a running commentary by French missionaries of the period 1830–1902; ELIZABETH A. ELDREDGE, *A South African Kingdom: The Pursuit of Security in Nineteenth-Century Lesotho* (1993); LEONARD THOMPSON, *Survival in Two Worlds: Moshoeshoe of Lesotho, 1786–1870* (1975); and PETER SANDERS, *Moshoeshoe, Chief of the Sotho* (1975), both analyzing Moshoeshoe's role in events; L.B.B.J. MACHOBANE, *Government and Change in Lesotho, 1800–1966: A Study of Political Institutions* (1990); and B.M. KHAKETLA, *Lesotho, 1970: An African Coup Under the Microscope* (1971). An excellent, comprehensive guide to the published material on Lesotho up to the time of publication is SHELAGH M. WILLET and DAVID P. AMBROSE, *Lesotho: A Comprehensive Bibliography* (1980). (J.J.Gu./Ed.)

Malaŵi. Overviews of the country can be found in HAROLD D. NELSON *et al., Area Handbook for Malawi* (1975, reprinted 1987). SWANZIE AGNEW and MICHAEL STUBBS (eds.), *Malawi in Maps* (1972); and MALAŴI DEPT. OF SURVEYS, *The National Atlas of Malaŵi* (1983?), present the country's physical characteristics and natural and human resources in cartographic form. MARGARET READ, *The Ngoni of Nyasaland* (1956, reissued 1970); and T. CULLEN YOUNG, *Notes on the History of the Tumbuka-Kamanga Peoples in the Northern Province of Nyasaland,* 2nd ed. (1970), are ethnographic studies. HORST DEQUIN, *Agricultural Development in Malawi* (1969), is a historical study of the period between 1890 and 1967. Economic conditions and politics are discussed in CAROLYN MCMASTER, *Malawi: Foreign Policy and Development* (1974); and T. DAVID WILLIAMS, *Malawi: The Politics of Despair* (1978).

Works chronicling the country's history are JOHN G. PIKE, *Malawi: A Political and Economic History* (1968); B.R. RAFAEL, *A Short History of Malawi,* 3rd ed. (1985); OWEN J.M. KALINGA, *A History of the Ngonde Kingdom of Malawi* (1985); BRIDGLAL PACHAI, *Malawi: The History of the Nation* (1973), and *Land and Politics in Malawi, 1875–1975* (1978); BRIDGLAL PACHAI (ed.), *The Early History of Malawi* (1972); JOHN MCCRACKEN, *Politics and Christianity in Malawi, 1875–1940: The Impact of the Livingstonia Mission in the Northern Province* (1977); RODERICK J. MACDONALD (ed.), *From Nyasaland to Malawi: Studies in Colonial History* (1975); IAN LINDEN and JANE LINDEN, *Catholics, Peasants, and Chewa Resistance in Nyasaland, 1889–1939* (1974); GEORGE SHEPPERSON and THOMAS PRICE, *Independent African: John Chilembwe and the Origins, Setting, and Significance of the Nyasaland Native Rising of 1915* (1958, reissued 1987); and PHILIP SHORT, *Banda* (1974).

(Z.D.K./K.M.G.P./J.C.Mi./K.In./Ed.)

Mozambique. HAROLD D. NELSON (ed.), *Mozambique, a Country Study,* 3rd ed. (1985), provides coverage of politics, security, economics, and society. REPÚBLICA POPULAR DE MOÇAMBIQUE, MINISTÉRIO DA EDUCAÇÃO, *Atlas Geográfico,* vol. 1, 2nd ed., rev. and updated (1986), is a useful geographic tool. *Country Profile: Mozambique* (annual), published by the Economist Intelligence Unit, contains accurate, up-to-date information on the economy, resources, and industry. STEPHANIE URDANG, *And Still They Dance: Women, War, and the Struggle for Change in Mozambique* (1989), highlights women's historical gains and their present struggles in the face of war and economic collapse. LINA MAGAIA, *Dumba Nengue: Run for Your Life: Peasant Tales of Tragedy in Mozambique* (1988; originally published in Portuguese, 1987), comprises a collection of testimony of the terror experienced by rural Mozambicans at the hands of Renamo. RUTH FIRST, *Black Gold: The Mozambican Miner, Proletarian, and Peasant* (1983), is a classic study of the country's historically most important flow of migrant labour from southern Mozambique to South Africa's gold mines. KEITH MIDDLEMAS, *Cabora Bassa: Engineering and Politics in Southern Africa* (1975), is a detailed study of the conceptualization, financing, and construction of one of the world's largest hydroelectric projects. *The Emergency Situation in Mozambique: Priority Requirements for the Period 1990–1991* (1990), prepared by the Mozambique government in collaboration with the United Nations, summarizes contemporary problems and development strategies.

Mozambique's history to independence is chronicled in MALYN NEWITT, *Portuguese Settlement on the Zambesi* (1973), and *A History of Mozambique* (1995); ALLEN F. ISAACMAN, *Mozambique: The Africanization of a European Institution: The Zambesi Prazos, 1750–1902* (1972); ALLEN F. ISAACMAN and BARBARA ISAACMAN, *The Tradition of Resistance in Mozambique* (1976), covering 1850–1921, and *Mozambique: From Colonialism to Revolution, 1900–1982* (1983); RENÉ PÉLISSIER, *Naissance du Mozambique: résistance et révoltes anticoloniales (1854–1918),* 2 vol. (1984), the most detailed source for the late slave era and conquest; LEROY VAIL and LANDEG WHITE, *Capitalism and Colonialism in Mozambique: A Study of Quelimane District* (1980), an excellent study of the colonial experience in Mozambique; THOMAS H. HENRIKSEN, *Revolution and Counterrevolution: Mozambique's War of Independence, 1964–1974* (1983), the most complete study in English of the independence struggle; and EDUARDO MONDLANE, *The Struggle for Mozambique* (1969, reissued 1983), a classic study of Mozambique's struggle to overcome colonial domination, written by the first president of Frelimo. Additional coverage of Mozambique can be found in the books by Birmingham and Minter, cited in the Angola section above. Works analyzing the revolution and subsequent events include JOSEPH HANLON, *Mozambique: The Revolution Under Fire* (1984, reprinted 1990), a sympathetic review of Frelimo's efforts to build a more egalitarian society, and *Mozambique: Who Calls the Shots?* (1991), a study of the country's economic and administrative near-collapse in the face of Renamo assaults, drought, and population displacement; ALEX VINES, *Renamo: Terrorism in Mozambique* (1991), a complete investigation of the organization's people, power structure, and international support networks; and WILLIAM FINNEGAN, *A Complicated War: The Harrowing of Mozambique* (1992), emphasizing the complex internal relationships which, along with the international dynamic of destabilization and military assaults in the country since independence, have fueled the destruction. COLIN DARCH and CALISTO PACHELEKE, *Mozambique* (1987), is an excellent annotated bibliography on all aspects of the country. (J.M.Pe./Ed.)

Namibia. BRIAN WOOD, *Namibia, 1884–1984: Readings on Namibia's History and Society* (1988), is a collection of somewhat uneven chapters on historical, social, and economic aspects. J.H. VAN DER MERWE (ed.), *National Atlas of South West Africa* (1983), is a detailed study of all aspects of Namibia's geography. RICHARD MOORSOM, "Underdevelopment, Contract Labor, and Worker Consciousness in Namibia, 1915–1972," *Journal of Southern African Studies,* 17:71–82 (October 1977), analyzes the nature, context, goals, and evolution of Namibian workers. UNITED NATIONS INSTITUTE FOR NAMIBIA, *Namibia: Perspectives for National Reconstruction and Development* (1986), brings together data and option analysis on most social, political, and economic aspects, although the policy advice is dated because it did not forecast reconciliation. DAVID SIMON and RICHARD MOORSOM, "Namibia's Political Economy: A Contemporary Perspective," in GERHARD TÖTEMEYER, VEZERA KANDETU, and WOLFGANG WERNER (eds.), *Namibia in Perspective* (1987), pp. 82–101, reviews the territorial economy, then approaching its low point. REGINALD H. GREEN, KIMMO KILJUNEN, and MARJA-LIISA KILJUNEN, *Namibia: The Last Colony* (1981), emphasizes the economy and the political-economic process. TORE LINNÉ ERIKSEN and RICHARD MOORSOM, *The Political Economy of Namibia: An Annotated Critical Bibliography,* 2nd ed. (1989), covers and comments on virtually all substantive material through 1988.

H. BLEY, *South-West Africa Under German Rule, 1894–1914* (1971; originally published in German, 1968), is a major study of the German occupation era with some coverage of the pre-colonial period. PETER H. KATJAVIVI, *A History of Resistance in Namibia* (1988), is a major study of Namibian history, especially from 1860 through the mid-1980s. REGINALD H. GREEN and P. MANNING, "Namibia: Preparations for Destabilization," in PHYLLIS JOHNSON and DAVID MARTIN, *Frontline Southern Africa: Destructive Engagement* (1988), pp. 153–189, gives an overview of the post-World War II liberation effort to the late 1980s. SWAPO DEPT. OF INFORMATION AND PUBLICITY, *To Be Born a Nation: The Liberation Struggle for Namibia* (1981), states the position, goals, and perception of history by the then-main nationalist movement, now the majority party. COLIN LEYS and JOHN S. SAUL, *Namibia's Liberation Struggle: The Two-Edged Sword* (1995), provides coverage of the recent liberation war and early years of independence. DAVID SIMON, *Independent Namibia: One Year On* (1991), is a review through early 1991; it may be supplemented by DONALD L. SPARKS and DECEMBER GREEN, *Namibia: The Nation After Independence* (1992). (R.H.Gr./Ed.)

South Africa. HAROLD D. NELSON (ed.), *South Africa, a Country Study* (1981), surveys South African society, economy, politics, and geography. *South African Review* (irregular); and *Race Relations Survey* (annual) document developments in politics, economy, and society. MONICA COLE, *South Africa,* 2nd ed. (1966), is a basic, comprehensive physical and human geography. SOUTH AFRICA DIRECTORATE OF SURVEYS AND MAPPING, *Reader's Digest Atlas of Southern Africa* (1984), despite its title, covers only South Africa in maps, photographs, and diagrams. An introduction to the social structure and politics of the country is BERNARD MAGUBANE, *The Political Economy of Race and Class in South Africa* (1978). DAVID M. SMITH (ed.), *The Apartheid City and Beyond* (1992); and MARK SWILLING, RICHARD HUMPHRIES, and KHEHLA SHUBANE (eds.), *Apartheid City in Transition* (1991), are collections of essays on urbanization, social change, urban politics, and development. Historical studies of the economy include ROBERT H. DAVIES, *Capital, State, and White Labour in South Africa, 1900–1960* (1979); and BELINDA BOZZOLI, *The Political Nature of a Ruling Class: Capital and Ideology in South Africa, 1890–1933* (1981). Contemporary economic conditions are treated in JILL NATTRASS, *The South African Economy: Its Growth and Change,* 2nd ed. (1988); NICOLI NATTRASS and ELISABETH ARDINGTON (eds.), *The Political Economy of South Africa* (1990); FRANCIS WILSON and MAMPHELA RAMPHELE, *Uprooting Poverty: The South African Challenge* (1989), a study of income distribution and aspects of poverty and unemployment; STEPHEN R. LEWIS, JR., *The Economics of Apartheid* (1990), an excellent review; and MERLE LIPTON and DAVID HAUCK (eds.), *The Impact of Sanctions on South Africa,* 2 vol. (1990). DEON GELDENHUYS, *The Diplomacy of Isolation: South African Foreign Policy Making* (1984), outlines government policies. (A.S.Ma.)

Broad coverage of South Africa's history is provided in MONICA WILSON and LEONARD THOMPSON (eds.), *The Oxford History of South Africa,* 2 vol. (1969–71), the only general reference work to make a serious attempt to record the history of all the peoples of South Africa; C.W. DE KIEWIET, *A History of South Africa, Social & Economic* (1941, reprinted 1978); LEONARD THOMPSON, *A History of South Africa* (1990), fluent and elegantly written; T.R.H. DAVENPORT, *South Africa: A Modern History,* 4th ed., updated and rev. (1991); and PAUL MAYLAM, *A History of the African People of South Africa: From the Early Iron Age to the 1970s* (1986), an overview for the nonspecialist of the history of African societies in South Africa. CHERRYL WALKER (ed.), *Women and Gender in Southern Africa to 1945* (1990), discusses the changing status of women in the past 100 years.

Early history to 1770 is explored in RICHARD ELPHICK, *Kraal and Castle* (1977; also published as *Khoikhoi and the Founding of White South Africa,* 1985), a detailed study of the interactions between the Dutch at the Cape of Good Hope and the Khoikhoi chiefdoms of the region; RICHARD ELPHICK and HERMANN GILIOMEE (eds.), *The Shaping of South African Society, 1652–1840* (1989), essays that review the history comprehensively from the first years of Dutch settlement; and S. DANIEL NEUMARK, *Economic Influences on the South African Frontier, 1652–1836* (1957). DAVID LEWIS-WILLIAMS and THOMAS DOWSON, *Images of Power: Understanding Bushman Rock Art* (1989), is an introduction to the rock paintings of southern Africa.

There is still no good published overview account of the period 1770–1870 in South African history, and there are still enormous gaps in knowledge. Nevertheless, the following titles are useful: CLIFTON C. CRAIS, *White Supremacy and Black Resistance in Pre-industrial South Africa: The Making of the Colonial Order in the Eastern Cape, 1770–1865* (1992); BEN MACLENNAN, *A Proper Degree of Terror: John Graham and the*

Cape's Eastern Frontier (1986), a study of the colonial invasion of the Zuurveld in 1811–12; W.M. MACMILLAN, *Bantu, Boer, and Briton: The Making of the South African Native Problem,* rev. and enlarged ed. (1963, reprinted 1978), a classic, still useful for the period 1820–40; PETER DELIUS, *The Land Belongs to Us: The Pedi Polity, the Boers, and the British in the Nineteenth-Century Transvaal* (1983); ROBERT ROSS, *Cape of Torments: Slavery and Resistance in South Africa* (1983); NIGEL WORDEN, *Slavery in Dutch South Africa* (1985); NOËL MOSTERT, *Frontiers: The Epic of South Africa's Creation and the Tragedy of the Xhosa People* (1992); JULIAN COBBING, "The Mfecane as Alibi: Thoughts on Dithakong and Mbolompo," *Journal of African History,* 29(3):487–519 (1988), arguing that the "Mfecane" is a species of colonial mythology; R. KENT RASMUSSEN, *Migrant Kingdom: Mzilikazi's Ndebele in South Africa* (1978), to the 1830s; J.B. PEIRES, *The House of Phalo: A History of the Xhosa People in the Days of Their Independence* (1981), up to the 1840s, and *The Dead Will Arise: Nongqawuse and the Great Xhosa Cattle-Killing Movement of 1856–7* (1989); COLIN BUNDY, *The Rise and Fall of the South African Peasantry,* 2nd ed. (1988), on the emergence of the African peasants after the 1840s; JEFF GUY, *The Destruction of the Zulu Kingdom: The Civil War in Zululand, 1879–1884* (1979), mainly on the British invasion of 1879 and its aftermath; and SHULA MARKS and ANTHONY ATMORE (eds.), *Economy and Society in Pre-industrial South Africa* (1980), essays on the pre-1900 period.

The period 1870–1930 is dealt with in H.J. SIMONS and R.E. SIMONS, *Class and Colour in South Africa, 1850–1950* (1969, reprinted 1983); WILLIAM H. WORGER, *South Africa's City of Diamonds: Mine Workers and Monopoly Capitalism in Kimberley, 1867–1895* (1987), a history of diamond mining, paying particular attention to questions of labour recruitment and the rise of the De Beers Mining Company; SHULA MARKS and RICHARD RATHBONE, *Industrialisation and Social Change in South Africa: African Class Formation, Culture, and Consciousness, 1870–1930* (1982), a collection of essays exploring some of the consequences of an industrial revolution for the country's African population; FREDERICK A. JOHNSTONE, *Class, Race, and Gold* (1976, reprinted 1987), an influential Marxist study of how racial discrimination was institutionalized in the gold-mining industry; T.R.H. DAVENPORT, *The Afrikaner Bond: The History of a South African Political Party, 1880–1911* (1966); CHARLES VAN ONSELEN, *Studies in the Social and Economic History of the Witwatersrand, 1886–1914,* 2 vol. (1982), a fine social history; THOMAS PAKENHAM, *The Boer War* (1979), a highly readable, well-researched popular history; PETER WARWICK (ed.), *The South African War* (1980); SHULA MARKS and STANLEY TRAPIDO, "Lord Milner and the South African State," *History Workshop,* 8:50–80 (Autumn 1979), an important article that led to the reevaluation of Milner's role in South Africa; LEONARD THOMPSON, *The Unification of South Africa, 1902–1910* (1960); DAVID YUDELMAN, *The Emergence of Modern South Africa: State, Capital, and the Incorporation of Organized Labor on the South African Gold Fields, 1902–1939* (1983); PETER WALSHE, *The Rise of African Nationalism in South Africa: The African National Congress, 1912–1952* (1970); HELEN BRADFORD, *A Taste of Freedom: The ICU in Rural South Africa, 1924–1930* (1987); GAIL M. GERHART, *Black Power in South Africa: The Evolution of an Ideology* (1978); and KEN LUCKHARDT and BRENDA WALL, *Organize or Starve!: The History of the South African Congress of Trade Unions* (1980).

South African history since 1930 is chronicled in numerous books. Recent syntheses of value include WILLIAM BEINART, *Twentieth-Century South Africa* (1994); and NIGEL WORDEN, *The Making of Modern South Africa: Conquest, Segregation, and Apartheid* (1994). JONATHAN CRUSH, ALAN JEEVES, and DAVID YUDELMAN, *South Africa's Labor Empire: A History of Black Migrancy to the Gold Mines* (1991), deals with the ways in which the history of the region has been connected through labour migrancy. FRANCIS WILSON, *Labour in the South African Gold Mines, 1911–1969* (1972), describes the dependence of South Africa's premier industry on African migrant workers and shows how the mining groups held down miners' wages so that they were lower in 1969 than in 1911. T. DUNBAR MOODIE, *The Rise of Afrikanerdom: Power, Apartheid, and the Afrikaner Civil Religion* (1975), shows how an Afrikaner civil religion, with antecedents going back to the 19th century, contributed to the political victory of the Afrikaner National Party in 1948. LEONARD THOMPSON, *The Political Mythology of Apartheid* (1985), shows with specific examples how the mythology of the Afrikaner nationalist movement was originally a mythology of liberation from British colonialism, but as British power waned it legitimated the oppression of the black people of South Africa. HERIBERT ADAM and HERMANN GILIOMEE, *Ethnic Power Mobilized: Can South Africa Change?* (1979), finds the roots of apartheid not in ideological racism or prejudiced Calvinism but in the entrenchment of Afrikaner power and privilege, mobilized against both black competitors and imperial foreign capital. DAN O'MEARA, *Volkskapitalisme: Class, Capital, and Ideology in the Development of Afrikaner Nationalism, 1934–1948* (1983), provides a useful account of a crucial period. MERLE LIPTON, *Capitalism and Apartheid: South Africa, 1910–84* (1985), explores, from a conservative perspective, the interaction of racial politics and economic interests in South Africa, especially in the years 1960–85, showing how capitalists who opposed apartheid became more influential after 1960. SHULA MARKS and STANLEY TRAPIDO (eds.), *The Politics of Race, Class, and Nationalism in Twentieth-century South Africa* (1987), collects essays by the "revisionist" or Marxist school of African historians emphasizing class conflict and capital accumulation. The essays in HERMANN GILIOMEE and LAWRENCE SCHLEMMER (eds.), *Up Against the Fences: Poverty, Passes, and Privilege in South Africa* (1985), describe the forces leading to the massive migration of Africans from the reserves to the cities and show that the government was failing to stop it. DAVID PALLISTER, SARAH STEWART, and IAN LEPPER, *South Africa Inc.: The Oppenheimer Empire,* rev. and updated ed. (1988), describes the global reach of the great industrial and financial conglomerate centred in the Anglo-American Corporation and the De Beers diamond cartel. JOSEPH HANLON, *Beggar Your Neighbours: Apartheid Power in Southern Africa* (1986), details how, during the 1980s, South African economic, political, and military power was used to destabilize other countries in southern Africa. WILLIAM MINTER, *King Solomon's Mines Revisited: Western Interests and the Burdened History of Southern Africa* (1986), offers a radical critique of the involvement of Britain, the United States, and other Western powers and financial interests in the exploitation of the black people of southern Africa. TOM LODGE, *Black Politics in South Africa Since 1945* (1983), is the basic history of black protest movements since World War II, with detailed examinations of specific campaigns and episodes. THOMAS KARIS and GWENDOLEN M. CARTER, *From Protest to Challenge: A Documentary History of African Politics in South Africa, 1882–1964,* 4 vol. (1972–77), is useful especially for its elaborate introductions; vol. 4 contains biographical sketches of major black politicians. SEBASTIAN MALLABY, *After Apartheid: The Future of South Africa* (1992), explores the options. DAVID OTTAWAY, *Chained Together: Mandela, De Klerk, and the Struggle to Remake South Africa* (1993), traces the common commitment of the white and black leaders to the transformation of South Africa. JACKLYN COCK, *Colonels & Cadres: War & Gender in South Africa* (1991; also published as *Women and War in South Africa,* 1993), explores the link between war and gender in South Africa in the final apartheid years. HERIBERT ADAM and KOGILA MOODLEY, *The Opening of the Apartheid Mind: Options for the New South Africa* (1993), provides a cogent analysis of the complex forces that operate in the new South Africa.

(M.Hl./J.R.D.C./C.J.B./L.M.T./Sh.M.)

Swaziland. The physical and human geography of the country are described in *Social Studies Atlas for Swaziland* (1991); G. MURDOCH, *Soils and Land Capability in Swaziland* (1968); ROBERT HAROLD COMPTON, *The Flora of Swaziland* (1976); BRIAN ALLAN MARWICK, *The Swazi* (1940, reissued 1966), an ethnographic account; HILDA KUPER, *An African Aristocracy: Rank Among the Swazi* (1947), on the social life and institutions of the Swazi; *Report on the 1986 Swaziland Population Census,* vol. 4, *Analytical Report* (1991?); D.C. FUNNELL, *Under the Shadow of Apartheid: Agrarian Transformation in Swaziland* (1991); CHRISTIAN P. POTHOLM, *Swaziland: The Dynamics of Political Modernization* (1972); and ALAN R. BOOTH, *Swaziland: Tradition and Change in a Southern African Kingdom* (1983).

Historical works include DAVID PRICE WILLIAMS, "Archaeology in Swaziland," *South African Archaeological Bulletin,* 35(131):13–18 (June 1980); PETER B. BEAUMONT, "The Ancient Pigment Mines of Southern Africa," *South African Journal of Science,* 69:140–146 (May 1973); J.R. MASSON, "Rock-paintings in Swaziland," *South African Archaeological Bulletin,* 16:(64):128–133 (December 1961); J.S.M. MATSEBULA, *A History of Swaziland,* 3rd ed. (1988); CAROLYN HAMILTON (ed.), *In Pursuit of Swaziland's Precolonial Past* (1990); PHILIP BONNER, *Kings, Commoners, and Concessionaires: The Evolution and Dissolution of the Nineteenth-Century Swazi State* (1983); and HILDA KUPER, *Sobhuza II, Ngwenyama and King of Swaziland: The Story of an Hereditary Ruler and His Country* (1978).

(J.R.M.)

Zambia. IRVING KAPLAN (ed.), *Zambia, a Country Study,* 3rd ed. (1979), provides introductions to all aspects of the country. D. HYWEL DAVIES (ed.), *Zambia in Maps* (1971), illustrates most aspects of the Zambian scene. Descriptions and maps of the traditional land use systems are found in JÜRGEN SCHULTZ, *Land Use in Zambia,* 2 vol. (1976). ROBERT E. BALDWIN, *Economic Development and Export Growth: A Study of Northern Rhodesia, 1920–1960* (1966), is a short but valuable analytical study by an economist, the closest approximation to a general economic history of Zambia. A.L. EPSTEIN, *Politics in an Urban African Community* (1958), is a celebrated study, both historical

and sociological, of the Roan Antelope mine compound and Luanshya township.

ANDREW ROBERTS, *A History of Zambia* (1976), is an overview. Colonial history is detailed in ROBERT I. ROTBERG, *Christian Missionaries and the Creation of Northern Rhodesia, 1880–1924* (1965); L.H. GANN, *The Birth of a Plural Society: The Development of Northern Rhodesia Under the British South Africa Company, 1894–1914*, 2nd ed. (1968, reprinted 1981); ANDREW ROBERTS, *A History of the Bemba: Political Growth and Change in North-Eastern Zambia Before 1900* (1973); and GWYN PRINS, *The Hidden Hippopotamus: Reappraisal in African History: The Early Colonial Experience in Western Zambia* (1980). A more specialized account of the colonial period can be found in BRIAN GARVEY, *Bembaland Church: Religion and Social Change in South Central Africa, 1891–1964* (1994). More recent events are discussed in ELENA L. BERGER, *Labour, Race, and Colonial Rule: The Copperbelt from 1924 to Independence* (1974); WILLIAM TORDOFF *et al.* (eds.), *Politics in Zambia* (1974); and MARCIA M. BURDETTE, *Zambia: Between Two Worlds* (1988), which reviews developments since independence with a focus on mining and the ailing economy. WILLIAM E. RAU, *A Bibliography of Pre-Independence Zambia: The Social Sciences* (1978), is a basic reference tool; it is complemented by GEOFFREY J. WILLIAMS, *Independent Zambia: A Bibliography of the Social Sciences, 1964–1979* (1984), covering the early years of independence. (G.J.Wi./A.D.R./Ed.)

Zimbabwe. Discussions of the country's geography, society, economy, and history are available in HAROLD D. NELSON, *Zimbabwe, a Country Study*, 2nd ed. (1983). Political economy is addressed by J.D.Y. PEEL and T.O. RANGER (eds.), *Past and Present in Zimbabwe* (1983); IBBO MANDAZA (ed.), *Zimbabwe: The Political Economy of Transition, 1980–1986* (1986); IAN PHIMISTER, *An Economic and Social History of Zimbabwe, 1890–1948: Capital Accumulation and Class Struggle* (1987); and CHRISTINE SYLVESTER, *Zimbabwe: The Terrain of Contradictory Development* (1991).

ROBERT BLAKE, *A History of Rhodesia* (1977), includes a commentary sympathetic to the white Rhodesian leaders. The early history of the country is detailed in D.N. BEACH, *The Shona & Zimbabwe, 900–1850: An Outline of Shona History* (1980); S.I.G. MUDENGE, *A Political History of Munhumutapa, c. 1400–1902* (1988); PHILIP MASON, *The Birth of a Dilemma: The Conquest and Settlement of Rhodesia* (1958, reprinted 1982), the best account of the early days (up to 1918) of white settlement and race relations; T.O. RANGER, *Revolt in Southern Rhodesia, 1896–97* (1967, reissued 1979), a full-length study, drawing from African sources, of the risings against white rule in 1896–97, with significance in terms of the modern liberation movement; ROBIN PALMER, *Land and Racial Domination in Rhodesia* (1977); ANTHONY VERRIER, *The Road to Zimbabwe, 1890–1980* (1986); T.O. RANGER, *The African Voice in Southern Rhodesia, 1898–1930* (1970); and CHARLES VAN ONSELEN, *Chibaro: African Mine Labour in Southern Rhodesia, 1900–1933* (1976), a major pioneering study in social history. LAWRENCE VAMBE, *An Ill-Fated People: Zimbabwe Before and After Rhodes* (1972), a family history portraying the humour and sadness of occupation, is continued by his *From Rhodesia to Zimbabwe* (1976), on the years from 1927 to the early 1960s. More recent history is studied by NATHAN M. SHAMUYARIRA, *Crisis in Rhodesia* (1965), a broad description of the racial disparities and political collisions that culminated in the Unilateral Declaration of Independence; RICHARD HALL, *The High Price of Principles: Kaunda and the White South* (1969); MARTIN MEREDITH, *The Past Is Another Country: Rhodesia, UDI to Zimbabwe*, rev. and extended ed. (1980), a detailed and objective account of political moves inside Rhodesia from 1965 to 1979; T.O. RANGER, *Peasant Consciousness and Guerrilla War in Zimbabwe: A Comparative Study* (1985); NORMA J. KRIGER, *Zimbabwe's Guerrilla War: Peasant Voices* (1992); DAVID MARTIN and PHYLLIS JOHNSON, *The Struggle for Zimbabwe: The Chimurenga War* (1981), an authoritative account of the liberation movement; and W.H. MORRIS-JONES (ed.), *From Rhodesia to Zimbabwe: Behind and Beyond Lancaster House* (1980).

(C.W.S./K.Br./K.In./Ed.)